

# The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line

Andrew Adey<sup>1,2</sup>, Joshua N. Burton<sup>1,2</sup>, Jacob O. Kitzman<sup>1,2</sup>, Joseph B. Hiatt<sup>1</sup>, Alexandra P. Lewis<sup>1</sup>, Beth K. Martin<sup>1</sup>, Ruolan Qiu<sup>1</sup>, Choli Lee<sup>1</sup>, Jay Shendure<sup>1</sup>

<sup>1</sup> Dept. of Genome Sciences, University of Washington, Seattle, WA 98115, USA

<sup>2</sup> These authors contributed equally to this work

**Correspondence should be addressed:**

Jay Shendure ([shendure@uw.edu](mailto:shendure@uw.edu))

Andrew Adey ([acadey@uw.edu](mailto:acadey@uw.edu))

# Supplementary Information

## Table of Contents

<b>Supplementary Notes</b>	<b>4</b>
S1   Shotgun sequencing and variant calling	4
S2   Indel calling with respect to coverage	4
S3   “SUNK” read depth determination	5
S4   Copy number state HMM	5
S5   Copy number recalibration and integer assignment	6
S6   Copy number outlier determination	6
S7   Copy number comparison of 8 additional HeLa strains	7
S8   Loss-Of-Heterozygosity (LOH) region calling	8
S9   Mate-pair library construction	9
S10   Identification of structural rearrangements	10
S11   Fosmid construction and sequencing	10
S12   Haplotype phasing	11
S13   Calling haplotype-resolved copy numbers (HRCNs)	13
S14   Ancestry-based analysis of HeLa haplotype phasing	14
S15   Haplotype analysis of 8 additional HeLa strains	15
S16   Identification of putative post-aneuploidy mutations	16
S17   HPV-18 integration site assembly	17
S18   Directional PolyA <sup>+</sup> RNA-Seq library construction	18
S19   RNA-Seq in-depth computational analysis	18
S20   ENCODE epigenome and RNA-Seq phasing	19
S21   Reference bias assessment and removal	19
S22   Identifying haplotype-specific peaks in ENCODE data	20
S23   Haplotype imbalance identification	20
<b>Supplementary Tables</b>	<b>21</b>
S1   Sequencing data obtained	21
S2   HGDP control genomes	23
S3   Summary of variants and regions of homozygosity for HeLa and control genomes	24
S4   Private protein-altering SNVs in HeLa CCL-2	(Excel)
S5   Private protein-altering indels in HeLa CCL-2	(Excel)
S6   HeLa CCL-2 private protein-altering SNVs/indels overlapping COSMIC or SCGC	25
S7   High-resolution copy number calls in HeLa CCL-2	(Excel)
S8   Haplotype-resolved copy number (HRCN)	27
S9   Large regions of LOH in HeLa CCL-2	28
S10   Structural rearrangements in HeLa CCL-2	(Excel)
S11   Genes in rearrangements in HeLa CCL-2	(Excel)
S12   Phasing status of heterozygous SNVs in HeLa CCL-2	29
S13   Haplotype-resolved copy numbers (HRCNs) in HeLa CCL-2	(Excel)
S14   Clone-confirmed somatic mutation frequency	30
S15   Variants shared between HeLa strains	31
<b>Supplementary Figures</b>	<b>32</b>
S1   Indel calling by coverage	32
S2   STR profiling with lobSTR	34
S3   Gene ontology enrichment analysis for genes with protein-altering variants in HeLa CCL-2 and 11 HGDP controls	35
S4   HeLa CCL-2 high resolution copy number calls	37
S5   HeLa over GC-matched control ratio histogram	38

S6	Copy-number recalibration strategy .....	39
S7	HeLa CCL-2 and S3 copy number and LOH profiles .....	41
S8	Mate pair insert size distributions .....	43
S9	Examples of deletions in HeLa CCL-2 .....	44
S10	Examples of inter-chromosomal rearrangements in HeLa CCL-2.....	45
S11	Called inversion examples in HeLa CCL-2.....	46
S12	Histogram of clone coverage .....	47
S13	Schematic of haplotype scaffolding approach using allele imbalance.....	48
S14	Gaussian mixture model of AAFs in non-LOH copy number 3 regions .....	49
S15	HeLa allele balance by read depth for HRCN regions.....	50
S16	Long read haplotype validation .....	51
S17	Allelic state across LOH event specific to HeLa S3 .....	52
S18	Population-based haplotype analysis .....	53
S19	Haplotype-based local inference of genetic ancestry.....	54
S20	Post-aneuploidy mutation analysis .....	57
S21	Somatic mutation allele frequencies .....	58
S22	Somatic mutation counts.....	59
S23	Private alleles shared between HeLa CCL-2 and S3 .....	60
S24	HeLa S3 high resolution copy number calls .....	61
S25	Copy number profiles for 10 HeLa strains .....	62
S26	Comparison of read depth profiles in HeLa strains .....	63
S27	Clustergram of 10 HeLa strains based on copy number profile similarity .....	64
S28	Regions of LOH in HeLa CCL-13 by comparison to CCL-2 haplotypes.....	65
S29	Copy-number loss and LOH on chromosome 9 in HeLa CCL-13.....	66
S30	Structure of the HPV-18 integration locus.....	67
S31	Assembly and sequencing of the HPV-18 integration site.....	71
S32	HPV-18 RNA-Seq coverage.....	73
S33	Correlation between RNA-Seq datasets.....	74
S34	Phased ENCODE data tracks .....	75
S35	Correlations between haplotype-specific signals for ENCODE HeLa datasets .....	76
S36	Reference bias in ENCODE peaks .....	77
S37	Minimal impact of reference bias upon transcript quantitation .....	78
S38	Reference bias removal.....	79
S39	Haplotype contributions of phased ENCODE data (windows).....	81
S40	Haplotype contributions of phased ENCODE data (box plots) .....	82
S41	Normalized haplotype imbalance scores by copy number.....	83
S42	Haplotype imbalanced ENCODE peak percentages .....	84
S43	ENCODE peak haplotype imbalance scoring .....	88
S44	ENCODE peak reference bias effects on outlier calling.....	89
S45	Phased ENCODE outlier analysis.....	91
S46	ENCODE haplotype imbalances for HPV-18 and MYC.....	93
S47	Long range with MYC from 5C and ChIA-PET data.....	94
S48	Datasets and analyses for HeLa CCL-2 and HeLa S3 .....	95

## S1 | Shotgun sequencing and variant calling

**Aim:** To construct shotgun sequencing libraries for all HeLa strains for use in variant calling, and copy number analysis.

**Input:** Cell cultures - HeLa: ATCC, CCL-2 (lab stock); HeLa S3 (lab stock): ATCC, CCL-2.2 (lab stock); Chang Liver: ATCC, CCL-13; L132: ATCC, CCL-5; KB: ATCC, CCL-17; HEp-2: ATCC, CCL-23; WISH: ATCC, CCL-25; Intestine 407: ATCC, CCL-6; FL: ATCC, CCL-62 and AV-3: ATCC, CCL-21.

All shotgun libraries were constructed using standard end-polishing, A-tailing, and ligation methods from 1 µg of genomic DNA (isolated using the QIAGEN Gentra PureGene kit). HeLa CCL-2 and HeLa S3 libraries were generated in duplicate at two different size ranges (2x 150-250 bp, and 2x 250-500 bp). All other HeLa isolate shotgun libraries were generated at a single size range (100-250 bp). All libraries were sequenced on an Illumina HiSeq 2000 with 13 lanes of paired-end 101 bp (PE101) reads for HeLa CCL-2 (six lanes required trimming read 2 to 40 bp due to an instrument solenoid valve failure on cycle 41 of read 2), 2 lanes of PE101 plus 2 lanes of PE51 for HeLa S3, and ½ lane of PE51 for all other HeLa isolates. For variant calling, reads were aligned to the human reference genome (hg19, b37) using BWA (v0.5.9)<sup>31</sup> with default parameters. Alignments of each library were merged and filtered to remove PCR duplicates, followed by quality score recalibration and indel realignment using the Genome Analysis ToolKit (GATK v1.6)<sup>32</sup>. SNVs were called using Samtools (v0.0.18)<sup>33</sup> due to its increased sensitivity for variants at low allele balances, and indels were called using GATK with a filter requiring a non-reference frequency of 0.1, coverage of at least 5, and quality score of at least 500.

## S2 | Indel calling with respect to coverage

**Aim:** To compare the total burden of indel variants between HeLa and control genomes, given effects from copy number and sequencing depth.

**Data Sources:** Shotgun indel calls from HeLa CCL-2 (~88X coverage), HeLa CCL-2 (downsampled to ~35X coverage), HeLa S3 (~26X coverage), and control genomes (~30-45X coverage), and corresponding shotgun bam files.

Markedly more indels were detected in HeLa than in HGDP control genomes analyzed in parallel (HeLa =  $4.2 \times 10^5$ , HGDP average =  $3.3 \times 10^5$ ). We hypothesized that this was simply due to increased coverage in HeLa than controls, so we downsampled the HeLa alignment to a comparable fold coverage (with respect to hg19) as the HGDP individuals (~35X) and re-called indels. This resulted in a markedly lower number of final calls ( $n = 2.1 \times 10^5$ ), partially reflecting the difference in ploidy between HeLa (aneuploid) and controls (diploid), resulting in lower coverage per copy number in HeLa. To investigate this further, we tallied indel and read depth in windows across the genome and plotted called indel counts with respect to sequencing depth, which revealed comparable trends for HeLa and controls (**Supplementary Fig. 1**).



### S3 | “SUNK” read depth determination

**Aim:** To establish coverage scores for windows in the genome of uniquely mappable positions termed “SUNK” windows (Singly Unique Nucleotide Kmer).

**Data Sources:** Sequence reads in fastq format (Samples and controls)

Reads were aligned using the mrsFAST read aligner<sup>36</sup>, which reports all possible alignments to a repeat-masked reference genome. Alignments were then processed by retaining unique alignments to non-overlapping windows each containing 50 uniquely mappable positions, or SUNK windows (singly unique nucleotide k-mers) as previously described<sup>37</sup>. These unique read counts per SUNK window were then used for read depth copy number analysis. (high-resolution: mean window size ~1.5 kb, 50 uniquely mappable positions per window) as well as for merged windows of 50 SUNK windows (low-resolution: mean window size: ~77 kb, 2,500 uniquely mappable positions per window).

### S4 | Copy number state HMM

**Aim:** To establish large-scale copy number states. Absolute copy number identification as well as identification of outlying copy number regions to be determined in later steps.

**Data sources:** SUNK read depth (Samples and controls)

SUNK window scores for HeLa CCL-2, HeLa S3, the 11 control genomes<sup>8</sup> and a GC-matched control library were first normalized to a constant to account for total read count differences. Scores for all samples were compared to one another to first identify windows of zero coverage, which may be population-specific deletions. These windows in HeLa were excluded from HMM copy number calling. Window ratio scores for HeLa strains over the GC-matched diploid normal control were then generated and served as the input observations to a basic HMM. These ratios were plotted as a histogram to identify approximate copy number ratio scores and increments between copy number states (**Supplementary Fig. 5**). Initiating ratios for copy number states were as follows: State ID = Ratio; 1 = 0.33, 2 = 0.66, 3 = 1.00, 4 = 1.33, 5 = 1.66, 6 = 2.00, 7 = 2.33, 8 = 2.66, 9 = 3.00, 10 = 3.33. These scores fit the histogram and also fit a triploid numerator, as might be expected given that the majority of the HeLa genome is at copy number three based on previous karyotypes in Macville *et al.* (1999)<sup>4</sup>. A maximum copy number of 10 was used, as nearly all of the HeLa genome falls below that copy number with the exception of small outliers that are detected in later steps. Emission scores for the HMM were determined by calculating the Gaussian probability for each of the state means using an initiating standard deviation of 0.1 for HeLa CCL-2 and 0.25 for HeLa S3:

$$E(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where  $x$  = the ratio score for the window,  $\mu$  = the mean ratio of each state, and  $\sigma$  = the standard deviation. Transition probabilities were initiated to 0.99999 within-state and 0.000001/ $N$  out of state (where  $N$  = the total number of other possible states). The low out-of-state transition probability was set so as to prevent overfitting manifested by transitions between copy number state between regions of different GC content. Additionally, each chromosome arm was segmented individually. One Viterbi iteration was deemed enough for convergence (as the initiating assumption of a triploid baseline, which was used in initializing the state means, is extremely close to the aggregate ploidy of HeLa).

## S5 | Copy number recalibration and integer assignment

**Aim:** To take into account aneuploidy to recalibrate copy number ratios and assign absolute copy number states.

**Data sources:** SUNK ratio scores segmented into states using an HMM.

To account for the effects of different sequencing depths, SUNK window scores were normalized to an equal value for each sample. This normalization makes the assumption that the case and control samples have equivalent masses of DNA in each cell, which is violated when comparing aneuploidy and diploid genomes (e.g., HeLa and control genomes). To account for this, a recalibration process was implemented that exhaustively assigns integer copy numbers to the states previously segmented by the HMM, without any preconceived assumptions about the relation between SUNK score and absolute copy number. These integers are then summed and divided by  $2N$  (where  $N$  is the total number of windows) to represent a diploid denominator resulting in a “Genetic Material Ratio” or GMR. The GMR is then applied as a recalibration constant to the SUNK windows and new state means are determined followed by comparison to the theoretical ratio scores expected under each copy number assignment hypothesis over a diploid control. The best-fitting hypothesis is then used and the states are assigned their respective absolute copy number values. A schematic of this process is shown in **Supplementary Fig. 6**.

## S6 | Copy number outlier determination

**Aim:** To identify outlier sets of windows that are not called during the HMM process due to the HMM’s stringent initial transition probabilities.

**Data sources:** Recalibrated SUNK ratio scores with copy number calls.

The HMM used to segment copy number states is designed to prevent overfitting that might occur due to local GC-content biases. As such, short copy number “outlier” regions were not segmented out and were detected separately. The outlier calling process utilized a sliding window approach that required 3 of 5 SUNK window ratio scores to be a minimum of 3 standard deviations away from the state mean in the same direction. Consecutive outlier windows were then merged, and a mean score for the outlying span was determined and assigned to an integer copy number based on theoretical ratio scores determined in the previous recalibration process (to a maximum copy number of 100). After outliers were called, standard deviations of each span were determined and if the standard deviation was  $> 0.5\mu$  ( $\mu$  = span mean), the window was split at various positions in which each new segment contained at least 5 SUNK windows. The split windows then had their mean and standard deviations calculated and were reassigned to a copy number. If the split provided a more optimum fit to copy number states over that of the original combined window then the split was retained. The splitting process was iterated until no splits were retained.

## S7 | Copy number comparison of 8 additional HeLa strains

**Aim:** To characterize the copy number profile of 8 additional HeLa strains and compare them to CCL-2 and S3.

**Data Sources:** Fastq files for shotgun sequencing libraries of 8 additional strains (Chang Liver: ATCC, CCL-13; L132: ATCC, CCL-5; KB: ATCC, CCL-17; HEp-2: ATCC, CCL-23; WISH: ATCC, CCL-25; Intestine 407: ATCC, CCL-6; FL: ATCC, CCL-62; AV-3: ATCC, CCL-21).

Reads were aligned and processed by two methods: 1) alignment followed by copy number calling as in **Supplementary Notes 4 and 5**, or 2) alignment with BWA (v0.5.9)<sup>31</sup>, followed by genotyping against all variants in HeLa CCL-2. For copy number profiling, integer copy number calls were estimated for all 8 strains as described for HeLa CCL-2 and S3 (**Supplementary Fig. 25**). The raw SUNK window scores (for low-resolution, 50 merged SUNK windows, ~77 kbp) were also compared in an all by all basis (**Supplementary Fig. 26**). These comparisons revealed that while all strains have HeLa copy number characteristics, they all have their own unique copy number profiles. Due to the batch-level differences in library construction and size selection between HeLa CCL-2 (Agarose size selection, 2 size ranges, 2 replicates each), HeLa S3 (PAGE size selection, 2 size ranges, 2 replicates each), and the additional 8 strains (PAGE size selection, 1 size range, 1 replicate), the direct comparisons between HeLa CCL-2 or HeLa S3 are notably noisier than between the additional 8 strains. In order to determine potential lineage information, large-scale windows were utilized (600 target SUNK windows, mean = 955,176 bp) and clustered using “pvclust” in R (**Supplementary Fig. 27**). This method utilized bootstrapping (we performed 1000 iterations) to assign confidence to the clustering dendrogram.

Shotgun reads from all 8 additional strains as well as from S3 (DNA-seq and RNA-Seq) were also aligned using BWA and genotyped at HeLa CCL-2 sites. Positions outside of segmental duplications, with coverage of at least 8X in both CCL-2 and the comparison strain were checked for the presence of the CCL-2 variant in at least one aligned read (**Supplementary Table 15**). Among whole-genome shotgun sequences from majority of strains, and RNA-seq from S3, 90-97% of interrogated positions showed evidence of the CCL-2 variant. The shared fraction was somewhat lower when only protein-altering variants were compared between CCL-2 and S3. To investigate this discrepancy, variants were split into bins of alternate allele coverage in CCL-2 and plotted for fraction of concordance. (**Supplementary Fig. 23**). In general, for variants supported by fewer reads in CCL-2, the fraction of sharing with S3 is reduced. This trend is observed for all private variants as well as among only private protein-altering variants with both approaching 100% sharing for alternate allele coverage in HeLa CCL-2 > 70X. Vertical bars represent a 95% binomial confidence interval. The decrease in concordance for protein-altering variants may be caused by sampling noise or sequencing artifacts enriched among those variants; however, the possibility of a true excess of strain-specific protein-altering variants cannot be conclusively ruled out.

## S8 | Loss-Of-Heterozygosity (LOH) region calling

**Aim:** To determine regions in the HeLa genome that have undergone loss of heterozygosity.

**Data sources:** SUNK window intervals; HeLa (CCL-2) shotgun SNV calls.

Loss of heterozygosity (LOH) was called using the local density of heterozygous markers by designing a hidden Markov model (HMM) with two states: presence and absence of LOH. At each interval used for copy number estimation (SUNK windows), the model emitted counts of homozygous and heterozygous variants called within that interval. To avoid falsely rejecting LOH due to the presence of spurious variants or somatic mutations, this analysis was restricted to SNVs from the shotgun data with VCF quality scores of at least 50 that overlapped with variants from the 1000 Genomes Project. Indels were excluded because their allele frequencies tended to be affected by reference mapping bias.

In each window, the quantity of homozygous and heterozygous SNVs was tallied, and the likelihood of observing this number of homozygous and heterozygous SNVs, given LOH or no LOH, was calculated (**Supplementary Fig. 7**). Each SNV was marked as being in or not in a repeat; SNVs were considered to be in repeat regions if they fell in any of the following UCSC Genome Browser tracks: “segdups.bed”, “repeat\_masked.bed”, “interrupted\_repeats.bed”, “simple\_repeats.bed”, “microsat\_repeats.bed”, and regions with wgEncodeCrgMapabilityAlign100mer  $\leq 0.05$ . It was assumed that SNVs in repeat regions would have a higher observed rate of heterozygosity due to mismapped reads. Specifically, in regions with LOH, the heterozygosity rate was assumed to be 0.5% of variants not in repeats and 10% of variants in repeats; while in regions without LOH, the heterozygosity rate was assumed to be 50% of variants not in repeats and 70% of variants in repeat. These round numbers were estimated from surveys of regions on chromosome 2 with clear signatures of LOH/non-LOH. The likelihood of each window’s set of observations (the number of homozygous/heterozygous variants in/not in repeats) was calculated as the binomial probability of those observations assuming the region was in LOH or not:

$$P(N_{het,r}, N_{hmz,r}, N_{het,nr}, N_{het,nr} | LOH) = (0.005)^{N_{het,nr}} \times (0.995)^{N_{hmz,nr}} \times (0.1)^{N_{het,r}} \times (0.9)^{N_{hmz,r}}$$

$$P(N_{het,r}, N_{hmz,r}, N_{het,nr}, N_{het,nr} | NO LOH) = (0.5)^{N_{het,nr}} \times (0.5)^{N_{hmz,nr}} \times (0.7)^{N_{het,r}} \times (0.3)^{N_{hmz,r}}$$

To guard against the risk of specious results in individual windows arising from phenomena such as mapping artifacts and sudden small-scale changes in copy number, the minimum possible value of P, for any SUNK window and LOH state, was set at  $10^{-6}$ . The HMM was initialized with transition probabilities of  $10^{-8}$  between the two states and equal initiation probabilities of the two states.

The HMM was run through a single iteration of Viterbi training, which was observed to be sufficient for convergence. The Viterbi training yielded a best path through the model, which indicated a prediction of the LOH state of each interval. Adjacent windows with LOH were merged, and the state of the most distal window on each chromosome arm was assumed to extend to the telomere. The resulting LOH calls are shown in **Supplementary Table 9**.

## S9 | Mate-pair library construction:

**Aim:** To construct mate-pair sequencing libraries of 3 kb and 40 kb.

**Input:** Genomic DNA (3 kb libraries), Pooled fosmid libraries (40 kb)

Library construction for 40-kilobase mate pair libraries was carried out similar to previously described methods in Gnerre *et al.* (2011)<sup>38</sup> starting with fosmid clone DNA pooled within each original fosmid preparation. Nicks introduced during clone DNA isolation were first repaired by incubating 10 µg pooled fosmid clone DNA with 10 units E. coli DNA Ligase I (NEB) in E. Coli DNA Ligase Reaction Buffer at 16°C for 60 min in a 50 µL reaction volume followed by denaturation at 65°C for 20 min and AMPure SPRI-bead cleanup (Agencourt). Introduction of nicks flanking the cloning site was performed by incubation at 37°C for 60 min with 25 units nicking restriction endonuclease Nb.BbvCI (NEB) in NEBuffer 2 and a 50 µL reaction volume and then denaturation at 80°C for 20 min followed by placing reactions on ice. Nicks were translated into the insert by addition of 10 units E. coli DNA Polymerase I (NEB) directly to the previous reaction followed by incubation for 45 min on ice and immediate denaturation at 80°C for 20 min then SPRI cleanup. Translated nicks were converted to double-stranded breaks by addition of T7 Exonuclease (NEB) in NEBuffer 4 and a 50 µL volume then incubation at 25°C for 2 hours, followed by SPRI cleanup and subsequent incubation with 0.2 µL S1 Nuclease (Fermentas) in S1 Nuclease Buffer at 25°C for 30 mins in a 30 µL reaction volume followed by addition of 2 µL 0.5M EDTA and heating to 80°C for 20 min. Fragments were end-repaired (NEB End-Repair Module) and 100 ng was then treated with 5 µL T4 DNA Ligase in a 500 µL reaction volume overnight at 25°C to promote intramolecular circularization. The circularized mate-pair fragments were amplified and converted to Illumina sequencing libraries by PCR with primers complementary to the vector backbone, followed by gel size selection. Libraries were pooled for sequencing with paired 100 bp reads on an Illumina HiSeq2000. Libraries of ~3 kilobase inserts were constructed following protocols described in Talkowski *et al.* (2011)<sup>39</sup>.

Reads were trimmed to 50 bp for the 40 kb libraries (to reduce the amount of reads which extend through the nick translation portion, through the ligation junction, and into the opposite segment) and to 25 bp for the 3 kb libraries (as the restriction enzyme utilized in the protocol cuts 25 bp away from the center ligation segment, thus only allowing 25 bp of genomic DNA), and then aligned to the human reference genome (hg19, b37) using BWA (v0.5.9)<sup>31</sup>, filtered for phred-scale mapping quality  $\geq 10$ , and filtered to remove PCR duplicates. For 40 kb libraries, a subset of read pairs with very short insert sizes (possibly resulting from translation of one but not both nicks) were suppressed, as were clusters of read pairs with nearly equal outer mapping coordinates. Insert size distributions can be found in **Supplementary Fig. 8**.

## S10 | Identification of structural rearrangements

**Aim:** Identification of deletions, inversions, and translocations from mate-pair data.

**Data Sources:** Aligned sequence reads from 3 kb and 40 kb mate-pair libraries.

Each alignment file was processed to sort potential rearrangement-spanning read pairs into one of five classifications: 1) concordant (expected read pair orientation, insert size: 3 kb:  $1,200 \text{ bp} \leq X < 5,000 \text{ bp}$ ; 40 kb:  $20,000 \text{ bp} \leq X < 60,000 \text{ bp}$ ), 2) short-pair (expected read pair orientation, insert size: 3 kb:  $X \leq 1,000 \text{ bp}$ ; 40 kb:  $X \leq 1,500 \text{ bp}$ ), 3) deletion (expected read pair orientation, insert size: 3 kb:  $X \geq 5,000 \text{ bp}$ ; 40 kb:  $X \geq 80,000 \text{ bp}$ ), 4) inversions (same chromosome, opposite read pair orientations, any insert size), and 5) translocations (different chromosome, any orientation).

Deletion, inversion, and translocation classifications for each library type were then subjected to a sliding window approach of 1 kb windows by 500 bp and the numbers of reads with start coordinates within each window were tallied. Windows with read counts in the top 5% had the respective read pairs investigated as to whether or not they fell into other top 5% ranking windows and the fraction of reads in the window falling into each of the other top 5% windows. For deletions and inversions a cutoff was set to make a call where at least 80% of the read pairs within the window link to the same alternate window, whereas translocations were set to require a 50% cutoff. Resulting calls were then merged to account for the overlapping nature of the sliding window approach followed by trimming the edges of the windows down to the first read pair identified that spans the link. Example calls can be found in **Supplementary Figs. 9-11**.

## S11 | Fosmid construction and sequencing

**Aim:** To construct and align fosmid dilution pools for purposes of haplotype phasing.

**Input:** HeLa (CCL-2) genomic DNA.

Three replicate fosmid libraries were prepared using the Epicentre CopyControl Fosmid Library Production Kit as previously described in Kitzman et al. (2011)<sup>5</sup> except for the use of a vector (GenBank Accession: EU140751.1) that was modified to contain Nb.BbvCI restriction enzyme sites at the ends of the vector to allow nicking to facilitate fosmid jumping library construction. Each fosmid library was then partitioned by limiting dilution into 96 sub-libraries which were then outgrown and converted into barcoded DNA-seq libraries by transposase-mediated tagging and fragmentation<sup>40</sup> and pooled for sequencing on a single lane of PE101 on a HiSeq 2000 for each fosmid set. Reads were aligned using BWA (v0.5.9)<sup>31</sup> with default parameters and filtering for a mapping quality phred score of 10.

## S12 | Haplotype phasing

**Aim:** To phase germline variants ascertained by shotgun sequencing onto haplotypes using fosmid clone pool sequencing and allele ratios.

**Data sources:** HeLa (CCL-2) whole-genome shotgun variants (SNV and indel calls); HeLa (CCL-2) fosmid clone dilution pool shotgun reads.

Deep whole-genome shotgun sequencing was used for discovery of SNVs and indels. Sub-genomic pools of long insert clones were sequenced in order to determine the haplotype phase for inherited (germline) variants. Because each pool sampled only a small fraction of the genome (median = 1697 clones/pool, or ~2.0% of the haploid genome, given a median insert size of 33 kbp), overlaps between clones within a given pool were expected to be rare, and the reads corresponding to a given clone could be assumed to derive entirely from one germline haplotype or another.

Clone inserts were mapped by a sliding window read depth approach, essentially as previously described in Kitzman *et al.* (2011)<sup>5</sup>. For each pool, reads with mapping quality  $\geq 20$  were counted within 1 kbp non-overlapping windows across the genome. Windows with low “mappability”, defined as those having fewer than 300 SUNKs (30mers unique within the genome), were excluded. A candidate clone location was recorded where, within a run of 20 to 45 consecutive mappable windows, at least 60% of the windows had read depth above the background level (defined as the 95<sup>th</sup> percentile of windowed read depths for the equivalent number of read mapping positions drawn at random from the genome). To map the boundaries of each clone insert, overlapping candidates were grouped, and the candidate was selected that maximized the score:

$$\begin{aligned} \text{candScore} = & \quad (\# \text{mappable windows in candidate interval with read depth} > \text{background}) \\ & - 2 \times (\# \text{flanking windows with read depth} > \text{background within } \pm 5 \text{ kbp}) \\ & + 10 \times (\% \text{ of mappable windows in candidate with read depth} > \text{background}) \end{aligned}$$

Regions with predicted LOH are haplotype-resolved with respect to germline variants by virtue of their hemizygosity, and were excluded from further analysis. Clones within non-LOH regions were intersected with heterozygous single-base and indel variants from the whole-genome shotgun data, limiting to likely germline variants (those found among individuals in the 1000 Genomes Project). Within each pool, variants at which clone-derived reads had discordant genotypes (indicative of overlapping clones or sequencing errors) were excluded, as were variants called in only a single read on the clone.

To merge individual clones into longer haplotype blocks, we used a custom implementation of the ReFHap algorithm<sup>41</sup>, which determines consensus haplotypes from the genotypes of overlapping haploid fragments (clones). Briefly, this algorithm creates a graph in which the nodes represent individual clones and the edges connecting them represent overlaps between the clones. Two clones  $c_a, c_b$  are considered to overlap if there are one or more variants covered by both clones. The edge representing this overlap is assigned a weight as follows:

$$W(c_a, c_b) = \sum_{\substack{\text{variants} \\ \text{called} \\ \text{by } c_a, c_b}} \begin{cases} 1 & \text{if calls are equal} \\ -1 & \text{if calls are unequal} \end{cases}$$

An estimated minimum cut is then calculated for the graph using iterations of the “GreedyInit” and “GreedyImprovement” steps, as in Duitama *et al.* (2010)<sup>41</sup>. The minimum cut determines the set of edges with the lowest possible total weight that can be removed in order to divide the graph into two disjoint subgraphs. The subgraphs represent the two germline haplotypes, and each clone was assigned to a

haplotype based upon which subgraph it was in, although determination of the major (“A”) and minor (“B”) haplotypes was not made until the scaffolding process, described below. Within each haplotype-representing subgraph, the clone phases were converted to variant phases. For each variant – including variants not present in 1000 Genomes – the set of all calls made by all clones at the variant was considered, taking the clones’ phases into account. If a majority of calls supported one phase over the other, that phase was taken as the correct phase of the variant; if there was an exact tie, the variant was not phased. For the vast majority of variants, the calls unanimously implied the same phase (**Supplementary Table 12**).

Next, haplotype blocks were further combined into longer “haplotype scaffolds” in regions of uneven copy number (*i.e.*, where one germline haplotype was present at greater copy than the other **Supplementary Fig. 13**). The principle is that, if a large genomic region has a consistent copy number such as 2:1 (*i.e.*, haplotype A is duplicated and haplotype B is not) then the variants from each haplotype should have distinct allele frequencies among the whole-genome shotgun reads – in this example, alleles on haplotype A should have shotgun read frequencies centered on 2/3 and alleles on haplotype B should have frequencies centered on 1/3. By convention, A and B are the haplotypes with more and fewer copies, respectively.

An HMM was used to combine all haplotype blocks within each contiguous interval of consistent, uneven copy number. The model contained three states: (1) haplotype A, (2) haplotype B, and (3) gaps between haplotype blocks. Each observation was a single variant that had been phased into haplotype blocks; each gap between adjacent haplotype blocks was also considered an observation. The transition probability between states A and B was initialized to  $10^{-8}$ , and transitions from the gap state into states A and B were equally likely (reflecting the lack of *a priori* knowledge of the relative phase of the haplotype blocks). At each variant, the HMM emitted the observed counts of whole-genome shotgun reads matching the alleles phased on each haplotype. The emission probability for each variant was then calculated as the likelihood of these read counts under a binomial distribution, parameterized by the total number of shotgun reads at each site and the predicted copy number of each haplotype:

$$P(\text{phase A} | \text{read counts, CNs}) = \left( \frac{CN_A}{CN_A + CN_B} \right)^{\text{counts}_A} \times \left( \frac{CN_B}{CN_A + CN_B} \right)^{\text{counts}_B}$$

The HMM was run through a single iteration of Viterbi training which was observed to be sufficient for convergence. The Viterbi training yielded a best path through the model, which indicated a prediction of the relative phase of all variants.

In addition to connecting adjacent haplotype blocks into longer scaffolds, these results assigned the “A” and “B” labels, indicating which haplotype was of higher copy. In addition, this model introduced switches within blocks between 0.19% of adjacent phased sites. These switches reflect likely errors on the part of fosmid-based phasing which were corrected by the signal of allelic imbalance among whole-genome shotgun reads.



### S13 | Calling Haplotype-Resolved Copy Numbers (HRCNs)

**Aim:** To create a genome-wide profile of HRCN – that is, the distinct copy numbers of each haplotype.

**Data sources:** HeLa (CCL-2) genome-wide copy-number and LOH calls; HeLa (CCL-2) haplotype scaffolds of phased variants.

HRCNs were called at all haplotype scaffolds, including unscaffolded blocks. First, the haplotype blocks/scaffolds were split at sites where the total copy number is predicted to change, and for each resulting interval, the most likely HRCN was determined according to the following process:

Haplotype-resolved copy numbers (HRCNs) are written here as [total copies]:[copies haplotype A]:[copies haplotype B]. For instance, regions at normal autosomal copy without LOH would be 2:1:1. Intervals coinciding with LOH regions were assigned an HRCN of  $CN_{total}:CN_{total}:0$ . Blocks/scaffolds in triploid regions without LOH are labeled as 3:2:1. Blocks/scaffolds in non-LOH regions of copy number 4 or more required special attention because more than one HRCN was possible; for instance, 4:3:1 and 4:2:2. For these cases, the alternate allele frequencies (AAFs) of all SNVs in the block/scaffold were tallied up, and each SNV's AAF was rounded to the nearest  $1/CN_{total}$ . For  $N$  in the range  $1...(CN_{total}-1)$ , each AAF rounded to  $N/CN_{total}$  was counted as evidence in favor of an  $CN_{total}:N:CN_{total}-N$  split (or  $CN_{total}:CN_{total}-N:N$ , if  $N < CN_{total}/2$ ) split. A value of  $N$  was called as the “correct” value if at least 10 SNVs, totaling at least 2/3 of the total number of SNVs, support it; otherwise, the evidence was considered inconclusive and no HRCN call was made. The total set of HRCN calls is shown in **Supplementary Table 8**. To interrogate potential selection on private protein-altering variants in HeLa CCL-2 we investigated the number of protein-altering variants that occur on the haplotype at a greater copy number than the wildtype. This resulted in 50.77% of private protein-altering SNVs and 43.64% of private protein-altering indels occurring on the amplified haplotype suggesting, at least globally, that there is no such correlation.

The breakpoints of HRCN spans were plotted as positions on chromosome ideograms to provide positional information that was used to generate blocks of contiguous haplotype of appropriate phase and copy number. This was then used along with marker chromosome descriptions from Macville *et al.* (1999)<sup>4</sup> as well as mate-pair structural calls to identify marker chromosomes and large-scale rearrangements likely present within the HeLa CCL-2 strain resulting in **Figure 1a**.

## S14 | Ancestry-based analysis of HeLa haplotype phasing

**Aim:** To analyze our scaffolds of phased variants by comparing them to expectations arising from the assumption of mixed European and West African ancestry in the HeLa genome.

**Data sources:** HeLa (CCL-2) haplotype scaffolds of phased variants; 1000 Genomes Project (CEU & YRI) variants with population frequencies.

The set of 60 CEU individuals and 59 YRI individuals from the 1000 Genomes Project was used as a reference panel. Each haplotype scaffold was partitioned into fixed windows of 1000 SNVs present in the reference panel, resulting in 1,161 windows across the genome. For both haplotypes on each window, a score  $S_{CEU}^H$  (net similarity of a haplotype to CEU) score was calculated, as the relative log-likelihood of the variants on that haplotype occurring in a CEU individual compared to a YRI individual:

$$S_{CEU}^H = \log_{10} \left[ \prod_{i=1}^{1000} \frac{f_{CEU}(v_i)}{f_{YRI}(v_i)} \right]$$

where  $f_{POP}(v_i)$  is the frequency of variant  $v_i$  in population  $POP$ . For variants appearing in one of the CEU, YRI populations but not the other, the frequency of the variant in the other population was set to a pseudocount value of 1/120 (*i.e.*, the equivalent of one occurrence in one haplotype among that population's reference panel.)

The values of  $S_{CEU}^H$  across all haplotypes in all windows form a clearly bimodal Gaussian distribution (**Supplementary Fig. 18a**). There is also a consistent negative correlation between  $S_{CEU}^H$  and the number of non-1000 Genomes Project variants present on a haplotype, consistent with a commonly observed enrichment of previously unknown variants on African haplotypes (**Supplementary Fig. 18b**). The haplotype blocks were called as either “CEU-like” or “YRI-like” based on thresholds of  $S_{CEU}^H \geq 0.1$  and  $S_{CEU}^H \leq -0.2$ , respectively. These calls allowed both haplotypes of the entire HeLa genome, outside of LOH regions, to be “painted” as either CEU-like or YRI-like (**Supplementary Fig. 19**).

## S15 | Haplotype analysis of 8 additional HeLa strains

**Aim:** To explore the haplotype patterns of the 8 HeLa strains sequenced to low coverage by comparing their haplotypes with that of HeLa CCL-2.

**Data sources:** HeLa (CCL-2) haplotype scaffolds of phased variants; HeLa (CCL-5,6,13,17,21,23,25,62) shotgun SNV calls with low read coverage.

The read coverage on the 8 HeLa strains is too low for *ab initio* LOH analysis, but it can be compared with the result of the variant phasing on HeLa strain CCL-2. The following analysis was performed on all 8 of the HeLa strains but is illustrated for CCL-13 only in **Supplementary Figs. 28 and 29**.

The genome was divided into large SUNK windows of ~800 kb each. In each window, the set of all variants phased in CCL-2 was inspected, and the coverage of these variant positions in reads from CCL-13 was tabulated. Each CCL-13 read covering one of these variant positions was phased as “A” or “B” according to the CCL-2 haplotype containing the allele seen in the CCL-13 read; then a total fraction of “A coverage” / “B coverage”, or “A/B ratio”, was calculated for the window.

**Supplementary Fig. 28** shows the A/B ratios for the entire genome (excluding windows in which HeLa CCL-2 is in LOH and thus contains no phased variants.) The A/B ratio is expected to be close to the allele balance of the CCL-2 haplotypes in CCL-13. Hence, in regions where the A/B ratio is close to 1 or 0, one of the CCL-2 haplotypes is likely absent in CCL-13. This can be seen in chr4q, chr9p, and chr18p. Also note regions in which the A/B ratio fluctuates rapidly between high and low numbers, such as chr4. These are areas in which CCL-2 has a balanced copy number, thus the haplotype blocks could not be combined into large-scale scaffolds, and the concept of haplotypes “A” and “B” is not expected to be consistent across the chromosome. **Supplementary Fig. 29** shows the A/B ratio and the copy number profile across chromosome 9 of HeLa CCL-13. CCL-13 has a similar copy number profile on chromosome 9 as CCL-2 (**Figure 1b**), except that the p arm is diploid rather than triploid. Notably, the A/B ratio is very close to 1 on all of chromosome 9p, implying that CCL-2 haplotype B is not present in CCL-13. These two lines of evidence strongly suggest that CCL-13 has lost its sole copy of the B haplotype of chromosome 9q, changing from a triploid 3:2:1 state to a diploid 2:2:0 state with LOH.

## S16 | Identification of putative post-aneuploidy mutations

**Aim:** To identify somatic mutations in HeLa that are very likely to have arisen after duplications in their region.

**Data sources:** HeLa (CCL-2) genome-wide LOH and haplotype-resolved copy number (HRCN) profiles; HeLa (CCL-2) haplotype scaffolds of phased variants.

We searched for candidate somatic mutations, starting from the set of all biallelic single-nucleotide variants ascertained by whole-genome shotgun sequencing ( $n=3,994,385$ ), followed by removal of homozygous sites (41.4% of sites with zero high-quality reads matching the reference allele). Of the remaining sites ( $n=2,339,608$ ), those present among the 1000 Genomes Project (86.7%) were taken as likely to be inherited rather than somatic variants, and were discarded. The remaining sites ( $n=311,847$ ) included true somatic mutations, rare or private inherited variants not observed within the 1000 Genomes Project dataset, and spurious calls corresponding to sequencing or mapping errors. To stringently exclude false positive mutations at sites of systematic error (e.g. artifacts occurring near repeat tracts or due to unannotated paralogs), we removed sites at which any one of the 11 control genomes' alignments contained the mutant allele (at a frequency of at least 10%), or where at least one of the control genomes had missing coverage (fewer than 10 reads). Additionally, we removed sites within annotated segmental duplications. After application of these filters, 66,829 sites remained.

We then selected sites that were polymorphic between duplicated copies of the same germline haplotype and are therefore *de facto* somatic mutations occurring after the haplotypes' duplication (**Supplementary Fig. 20**). To find such sites, we searched dilution pools for clones derived from the same germline haplotype but with differing genotypes at the candidate site. Clones derived from the opposite germline haplotype were required to match the reference allele (except for in LOH regions, where only one germline haplotype remains). We excluded sites at which haplotype-resolved copy number was ambiguous (either uncalled, or where calls based upon low- and high-resolution windows were discordant), as well as sites lacking coverage from any phased clones. In sum, this confirmed 8,165 somatic mutations (**Supplementary Table 14**), throughout the genome (**Supplementary Fig. 22**)

Requiring observation of both the mutant and wild-type alleles on distinct clones from the same germline haplotype is a stringent filter, but it rejects true mutations in cases where not enough clones are sampled to observe both the mutant and wild-type alleles on the same germline haplotype. To estimate the proportion of true somatic mutation sites lost by undersampling, we considered the expected number of cases in which our method would fail to sample enough clones to observe both alleles among sites containing true variants (sites shared between the HeLa genome and the 1000 Genomes Project). We started with 1,203,938 phased, heterozygous sites present within HRCN=3:2:1 regions. At the 21.5% of sites covered by zero or one clones from the duplicated haplotype, it would have been impossible to observe both alleles in separate clones. For sites with at least two clones from the duplicated haplotype, we computed the expected number of sites for which both wild-type and mutant clones would have been observed under the assumption that each are sampled with equal likelihood:

$$n_j = [\text{\#sites with } j \text{ clones derived from duplicated haplotype}]$$

The expected number of sites with  $j$  clones derived from duplicated haplotype, where at least one wild-type and one mutant clone are observed is:

$$x_j = n_j (1 - 0.5^{j-1})$$

The overall expected sensitivity is:

$$\frac{\sum_{j=2}^{\infty} x_j}{\sum_{j=0}^{\infty} n_j}$$

By this measure, the expected sensitivity for validation of somatic sites discovered by shotgun sequencing is 0.611. Extrapolating from HRCN=3:2:1 regions provides a conservative (low) estimate for sensitivity, because clone sampling depth is greater within more heavily duplicated regions (that is, those for which the duplicated haplotype is present at more than two copies), increasing the likelihood of sampling enough clones to observing both alleles. Therefore, the expected true number of somatic mutations is no more than  $[8,165/(0.611)] = 13,364$ .

For a large majority of the clone-confirmed somatic mutations, the shotgun allele frequencies (**Supplementary Fig. 21**) are consistent with the idea that the mutations occurred after all duplications have taken place and the cell line has reached stable copy, although these allele frequencies were not used as a criterion for selecting the mutations (other than to exclude invariant sites - those with allele frequency equal to zero or one). Among regions with HRCN=3:2:1, 3,919 sites (94.9%) and among 3:3:0 regions, 1,334 sites (92.8%) had shotgun allele frequencies less than 50%, over which range presence on only one copy of duplicated haplotype A is more likely than presence among both. Among HRCN 4:2:2, 4:3:1, and 4:4:0 regions, the respective counts and proportions of sites consistent with mutation strictly after duplication, and therefore presence on only one copy (that is, sites with allele frequency  $\leq 37.5\%$ ), were 231 (80.5%), 153 (70.8%), and 809 (90.8%). The 3% of the genome with higher copy number is omitted from this analysis, because of the potential for bias against the variants at low shotgun allele frequencies.

## S17 | HPV-18 integration site assembly

**Aim:** To assemble the complex repetitive structure of the HPV-18 integration site on chromosome 8q24.21.

**Data sources:** Fosmid clone dilution pool reads; Final copy number calls; Raw shotgun reads and read depth; PCR assays.

In order to first identify potential locations of HPV-18 integration in the genome, the whole-genome shotgun and fosmid clone dilution pool reads were aligned to a modified reference that contained the HPV-18 genome sequence as well as the sequence of the fosmid vector backbone (for clone end determination). Clone pools were then sorted based on coverage with respect to the HPV-18 genome and read pairs flagged as interchromosomal between HPV-18 and a human chromosome were used to determine the site of integration in the human reference sequence. This region was constrained exclusively to chromosome 8q24.21 and therefore all pools with coverage spanning chromosome 8q24.21 were also pulled for further analysis. Potential breakpoints were then determined using the breakpoint-spanning read pairs in the fosmid pools as well as the shotgun read pairs. These breakpoints were then confirmed by breakpoint PCR of all possible primer pair combinations for each breakpoint, followed by the construction of shotgun libraries of the amplicons using transposase-based library preparation<sup>40</sup> and then sequencing on a MiSeq (**Supplementary Fig. 31**). Coverage profiles were then generated for clone pools with coverage of HPV-18 and/or chromosome 8q24.21 (**Supplementary Fig. 30**). These coverage profiles of a fixed expected length were then used in conjunction with shotgun read depth as well as copy number calls to determine the exact repetitive structure of the integration locus (**Figure 2**).

## S18 | Directional PolyA<sup>+</sup> RNA-Seq library construction

**Aim:** To generate high-depth in-house directional, PolyA RNA-Seq libraries in order to better assess effects of copy number and transcription.

**Input:** HeLa S3 cell culture.

HeLa S3 was chosen over HeLa CCL-2 due to the clonality of the strain. However, copy number heterogeneity has been observed in HeLa S3, it is notably less than that of HeLa CCL-2. Total RNA was isolated using the RNeasy mini kit (QIAGEN) followed by quantification using a NanoDrop 8000 spectrometer. 1 µg of total RNA was then used for mRNA isolation using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) followed by directional RNA-Seq library preparation using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB). All kit protocols were carried out according to manufacturer instructions. The library was then sequenced on one lane of HiSeq 2000 using paired-end 51 bp reads.

## S19 | RNA-Seq in-depth computational analysis

**Aim:** To provide additional RNA-Seq depth of the HeLa S3 strain.

**Data Sources:** HeLa S3 RNA-Seq fastq files (from Nagaraj *et al.* (2011), ENCODE CSHL Long PolyA Cell, and in-house directional RNA-Seq).

Reads were aligned to hg19 as well as haplotype-specific HeLa reference for haplotype A and haplotype B (**Supplementary Note 21**) with TopHat (v2.0.6)<sup>42</sup> using the RefGene “gtf” file downloaded from the UCSC genome browser. The hg19 alignment was then used for transcript quantification using Cufflinks (v2.0.2)<sup>43</sup>. For the in-house RNA-Seq, quantification was also performed on each haplotype alignment resulting in minimal differences (**Supplementary Note 21, Supplementary Fig. 37**). Transcripts with RPKM scores greater than or equal to 1 were used for further comparisons to mitigate noise associated with inactive transcripts. Correlations were performed using the “cor.test” function in R for both Pearson and Spearman tests. Additionally, a correlation between in-house RNA-Seq and ENCODE (CSHL, Cell, Long, PolyA) RNA-Seq was also performed (Spearman = 0.646, Pearson, 0.363; **Supplementary Fig. 33**). Lastly, reads unaligned to the human genome were aligned to the HPV-18 reference to investigate transcription levels of the integrated viral genome (**Supplementary Fig. 32**).

Global transcription levels by copy number were assessed by using genes in regions of constant copy number between CCL-2 and S3 and split by underlying copy number. This resulted in an increasing trend with a *p*-value of 0.075 according to a permutation analysis by which copy number identities are shuffled at each iteration (to a total of 100,000 iterations), yet retaining the total number of genes in each copy number bin (**Figure 3a**). Scores were then normalized to underlying copy number and the test performed again which resulted in a *p*-value of 0.485 according to the previously described permutation analysis (**Figure 3b**). While the increasing trend is not significant enough to definitively claim increased expression by copy number, the comparison to the copy number normalized *p*-value which is extremely near the null hypothesis is convincing nonetheless.

## S20 | ENCODE epigenome and RNA-Seq phasing

**Aim:** To assign haplotype phase to transcripts and epigenomic data tracks generated on HeLa.

**Data Sources:** In-house generated directional, PolyA RNA-Seq on HeLa (CCL-2) and HeLa (S3) as aligned bam files, downloaded HeLa (CCL-2) RNA-Seq from Nagaraj *et al.* (2011)<sup>6</sup> as aligned bam files, and downloaded ENCODE data sets on HeLa (S3) comprised of: DNase (UW), FAIRE-Seq (UNC, 2 tracks), Histone modifications (Broad, 13 tracks), Histone modifications (UW, 3 tracks), Repli-Seq (UW, 6 tracks), RNA-Seq (CSHL, 9 tracks), RNA-Seq (CalTech, 4 tracks), Transcription factor binding (Hudson-Alpha, 4 tracks), Transcription factor binding (Stanford, Yale, Duke, Harvard, 48 tracks) as aligned bam files and with called peaks where appropriate.

For all RNA-Seq, RPKM (Reads Per Kilobase per Million) scores for global levels of transcription were generated by tallying the number of reads per kilobase window of the genome. RPKM scores were also generated by a gene-model based approach using Cufflinks (v2.0.2)<sup>43</sup>. All bam files were then genotyped for all phased variants, and the fractional contribution of each haplotype to each RPKM or peak score was calculated. Copy number normalization was then performed by dividing the haplotype-specific score by the underlying copy number of that haplotype to find the haplotype contribution per copy. A global view of these tracks can be found in **Supplementary Fig. 34**.

## S21 | Reference bias assessment and removal

**Aim:** To identify the contribution of reference bias in alignment and subsequent allele balance calculation.

**Data Sources:** Raw fastq sequence read files, homozygous and phased heterozygous SNVs in HeLa, peaks called for corresponding sequence tracks.

Reference bias in allele balance at informative sites was determined by calculating the haplotype A / (haplotype A + haplotype B) ratio at each HRCN classification split by sites where the reference allele is either haplotype A or haplotype B. This resulted in globally lower ratios for positions where haplotype B is the reference allele and is summarized for all regulatory peaks in HRCN 3:2:1 regions in **Supplementary Figs. 36 and 38**. To remove this bias, two new reference genomes were generated with all homozygous SNVs as well as either haplotype A or haplotype B heterozygous SNVs (referred to as HAPREF) followed by alignment of raw reads for a subset of the ENCODE data sets as well as in-house RNA-Seq to each reference and tallying counts for respective heterozygous calls for each haplotype. This process effectively removed the reference bias, as shown in **Supplementary Fig. 38**. We next assessed the effect of reference bias on haplotype-specific peak calling (**Supplementary Note 22**) by comparing results from the hg19 and HAPREF data sets revealing only a very minor shift in the number of called outliers (**Supplementary Fig. 44**). For the in-house RNA-Seq of HeLa S3, gene transcript quantifications were made on the hg19 alignment as well as for each individual haplotype reference and compared (**Supplementary Fig. 37**) resulting in 0.54% of transcripts with a difference in RPKM score  $\geq 10\%$  between haplotype A and haplotype B references with 0.63% and 0.64% of transcripts showing a  $\geq 10\%$  difference for haplotype A and haplotype B references respectively when compared to hg19.

## S22 | Identifying haplotype-specific peaks in ENCODE data

**Aim:** To identify peaks originating almost exclusively from a single haplotype.

**Data Sources:** Phase-resolved peaks from epigenetic data tracks.

In order to quantitatively assess the number of haplotype specific peaks for ENCODE data tracks, a scoring metric was derived that takes into account both the statistical significance of the peak allele balance differing from the null hypothesis of the HRCN theoretical allele balance as well as quantifying the bias of the allele balance away from the null hypothesis of the HRCN mean allele balance. The first score was calculated using the following probability mass function:

$$f(t; a; c_A; c_B) = \left( \frac{t!}{a! (t-a)!} \right) \left( \frac{c_A}{c_B} \right)^a \left( 1 - \frac{c_A}{c_B} \right)^{t-a}$$

where  $t$  = the total coverage at the position,  $a$  = the number of bases supporting the haplotype A allele, and  $c_{A/B}$  = the copy number for haplotype A or haplotype B at the position. The resulting score is a  $p$ -value corresponding to the significance that the alleles observed at the peak are different from the null hypothesis. This metric only provides a  $p$ -value against the null hypothesis. In order to quantify the difference in allele balance, a second normalized Gaussian score was applied:

$$f(x) = \left( \frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where  $x$  = the haplotype A / (haplotype A + haplotype B) ratio at the position,  $\sigma$  = the standard deviation of haplotype A / (haplotype A + haplotype B) ratios for the entire HRCN state, and  $\mu$  = the mean of haplotype A / (haplotype A + haplotype B) ratios for the entire HRCN state. This score (shown for CTCF binding peaks in **Supplementary Fig. 41**) is then normalized across all HRCN classifications. These two scores can then be used with a cutoff to identify peaks of extreme haplotype imbalance (**Supplementary Figs. 42 and 43**).

## S23 | Haplotype imbalance identification

**Aim:** To identify regions of excessive haplotype imbalance.

**Data Sources:** Phased epigenomic and transcriptomic data sets.

Regions of excessive haplotype imbalance were identified using a sliding window approach (1.5 Mb sliding by 0.5 Mb) which took into account the ranking of the peak weight within each respective data set with a cap set at the top 50th percentile (i.e. smallest score of 0.5 and decreasing out to 1.0 as the smallest peak, so as not to over-weight excessively high peaks) as well as the haplotype imbalance score for the dominating haplotype (positive score for A, negative score for B) based on coverage at haplotype-resolved heterozygous variants and normalized for their underlying haplotype copy number which was set to a maximum ratio of 10 for haplotype A and -10 for haplotype B in order to minimize noise. The score for each track was calculated by dividing the haplotype imbalance score by the peak weighting score to produce maximums of 20 for haplotype A and -20 for haplotype B. The absolute values for all tracks within the window were then summed to produce the final window score. Several iterations of the capping values were implemented to highly similar results; these final values were used as they mitigated noise caused by single, dominating peaks and tended to favor windows containing multiple haplotype imbalanced data sources. Windows were then ranked and filtered to remove regions of LOH followed by combining like-windows in the top set of hits and rescored merged windows (**Supplementary Fig. 45**).



## a Shotgun Sequencing

ID	Unique Read Pairs	Unique reads (%)	Insert Size (bp)	Aligned Bases (Gbp)	Fold Coverage*
HELA					
HELA.s.1	297,931,188	97.5	142 +/- 29	48.36	17.27
HELA.s.2	155,257,336	96.3	131 +/- 28	32.80	11.71
HELA.s.3	527,420,302	90.5	206 +/- 39	94.32	33.69
HELA.s.4	430,404,271	85.3	196 +/- 45	72.28	25.81
<b>TOTAL</b>				<b>247.76</b>	<b>88.48</b>
HELA S3					
S3.s.1	241,974,865	93.8	267 +/- 119	29.15	10.41
S3.s.2	234,719,279	94.2	270 +/- 116	28.09	10.03
S3.s.3	32,257,707	97.1	289 +/- 122	6.62	2.36
S3.s.4	42,792,269	98.2	300 +/- 114	8.76	3.13
<b>TOTAL</b>				<b>72.62</b>	<b>25.93</b>

## b Fosmid Clone Pool Sequencing

ID	Called Clones	Average Clone Size (bp)	Physical Coverage*
Hapfos1	171,580	33,495	2.05
Hapfos2	228,667	34,851	2.85
Hapfos3	118,046	33,204	1.40
<b>TOTAL</b>	<b>518,293</b>	<b>-</b>	<b>6.30</b>

## c

### Mate Pair Sequencing

ID	Type	Unique Concordant Pairs	Insert Size (bp)	Physical Coverage*	Unique Discordant Pairs (Intra)	Unique Discordant Pairs (Inter)	Unique Discordant Pairs (Inversion)
Matepair1	circularization	85,311,942	2,862 +/- 453	87.20	356,696	10,522,598	300,785
Matepair2	circularization	46,363,211	2,861 +/- 453	47.38	209,998	6,041,976	178,559
<b>TOTAL</b>				<b>134.58</b>			
Matefos1	fosmid end	86,462	34,992 +/- 4,309	1.08	248	3,279	207
Matefos2	fosmid end	163,115	35,969 +/- 4,211	2.10	321	4,895	400
Matefos3	fosmid end	94,503	35,064 +/- 3,321	1.18	6,367**	94,588**	6,471**
<b>TOTAL</b>				<b>4.36</b>			

## d RNA-Seq

ID	Unique Read Pairs	Insert Size (bp)	Correct Strand (%)	Ribosomal (%)	Aligned to Coding (%)	Aligned to UTR (%)
S3.RNA	227,472,084	173 +/- 46	99.42	0.07	45.68	41.73

\*Assuming 2.8 Gbp alignable reference

\*\*Higher discordant rate due to increased intermolecular ligation events

## e HeLa 8 Strains

ID	Name	Total Aligned Bases (Gb)	Insert Size (bp)	Coverage*
CCL-13	Chang Liver	11.3	138 +/- 100	4.0
CCL-5	L132	12.1	138 +/- 109	4.3
CCL-17	KB	10.5	145 +/- 110	3.8
CCL-23	HEp-2	10.2	147 +/- 107	3.7
CCL-25	WISH	10.1	138 +/- 109	3.6
CCL-6	Intestine 407	10.7	140 +/- 70	3.8
CCL-62	FL	10.8	146 +/- 95	3.9
CCL-21	AV-3	9.8	144 +/- 107	3.5

## f PacBio RS Long Read Sequencing

ID	Total Reads	Aligned Reads	Aligned Informative (inf) Reads**	Aligned Read Length (all, bp)	Aligned Read Length (inf, bp)
HELA.PB5KB	601,217	114,584	6,746	1,428 +/- 1488	2970 +/- 1645

\*Assuming 2.8 Gbp alignable reference

\*\* Reads overlapping at least 2 heterozygous, phased SNVs with aligned positions  $\geq 10$ bp from nearest alignment indel

### Table S1 | Sequencing data obtained.

Six major types of sequence data were obtained. **a.** Shotgun sequencing data obtained for HeLa (CCL-2) and HeLa S3. **b.** Haplotype specific fosmid clone pool sequencing. **c.** Mate pair sequencing for both 3kb jumping libraries as well as fosmid based 40kb jumping libraries. **d.** Directional PolyA RNA-Seq library for HeLa S3 generated in house. **e.** Shotgun sequencing of 8 additional HeLa strains. **f.** PacBio RS long read sequencing of HeLa CCL-2 for haplotype phasing validation.

	ID	iSize Mean	iSize Stdev	Total Aligned Bases (Gb)	Fold Coverage*
Dinka	DNK02	258	117	92.37	32.99
French	HGDP00521	283	124	95.35	34.05
Papuan	HGDP00542	260	119	92.83	33.15
Sardinian	HGDP00665	270	119	88.48	31.60
Han	HGDP00778	269	120	97.74	34.91
Yoruban	HGDP00927	279	120	113.82	40.65
Karitiana	HGDP00998	267	128	96.69	34.53
San	HGDP01029	272	126	122.75	43.84
Madenka	HGDP01284	264	123	91.14	32.55
Dai	HGDP01307	256	125	97.40	34.79
Mbuti	HGDP0456	265	120	85.89	30.68

\*Assuming 2.8 Gbp alignable reference

**Table S2 | HGDP control genomes.**

Sequencing data summary for 11 HGDP control individuals from Meyer M. *et. al.* (2012).

	African	European		European		African
	DNK02, Dinka	HGDP00521, French	HGDP00542, Papuan	HGDP00665, Sardinian	HGDP00778, Han	HGDP00927, Yoruban
Number of SNVs	4,490,051	3,787,833	3,722,767	3,777,671	3,833,823	4,588,782
Number of indels	359,565	326,732	301,328	314,531	324,555	378,632
Number of 1kG SNVs	4,014,503	3,403,243	3,192,306	3,383,120	3,420,873	4,128,798
Number of non-1kG SNVs	475,548	384,590	530,461	394,551	412,950	459,984
Number of 1kG indels	213,292	183,321	170,851	179,872	183,241	221,893
Number of non-1kG indels	146,273	143,411	130,477	134,659	141,314	156,739
% SNVs that are homozygous	35.38%	39.33%	49.19%	39.57%	42.36%	35.06%
Ti/Tv for SNVs in 1kG	2.15	2.15	2.15	2.15	2.14	2.15
Ti/Tv for SNVs not in 1kG	1.62	1.59	1.70	1.61	1.58	1.59
Private Protein-Altering (PPA) SNVs	304	117	678	199	249	143
PPA SNVs in COSMIC	3	0	7	1	2	1
PPA SNVs in Cancer Genes	10	5	17	3	4	4
PPA indels	8	8	16	13	19	19
PPA indels in COSMIC	0	0	0	0	0	0
PPA indels in Cancer Genes	0	0	0	0	0	0
Total bases in homozygous tracts	30,173,695	42,933,724	124,265,489	62,502,510	50,148,051	62,624,662

	African		African		African
	HGDP00998, Karitiana	HGDP01029, San	HGDP01284, Madenka	HGDP01307, Dai	HGDP0456, Mbuti
Number of SNVs	3,570,099	5,015,502	4,621,804	3,774,107	4,783,268
Number of indels	303,842	363,202	352,080	321,657	337,610
Number of 1kG SNVs	3,156,213	4,009,840	4,108,845	3,380,387	4,100,827
Number of non-1kG SNVs	413,886	1,005,662	512,959	393,720	682,441
Number of 1kG indels	169,091	206,717	214,807	180,726	204,927
Number of non-1kG indels	134,751	156,485	137,273	140,931	132,683
% SNVs that are homozygous	51.21%	37.33%	34.56%	42.46%	38.21%
Ti/Tv for SNVs in 1kG	2.14	2.15	2.15	2.14	2.15
Ti/Tv for SNVs not in 1kG	1.62	1.81	1.65	1.59	1.76
Private Protein-Altering (PPA) SNVs	276	1258	212	238	626
PPA SNVs in COSMIC	1	7	3	2	2
PPA SNVs in Cancer Genes	10	18	5	5	15
PPA indels	7	48	14	6	35
PPA indels in COSMIC	0	0	0	0	0
PPA indels in Cancer Genes	1	0	0	0	0
Total bases in homozygous tracts	327,170,977	51,559,155	25,223,209	65,259,057	77,220,484

	AVERAGES			
	5 African	2 Euro	Reich 11	HELA CCL-2
Number of SNVs	4,699,881	3,782,752	4,178,701	4,068,395
Number of indels	358,218	320,632	334,885	417,471*
Number of 1kG SNVs	4,072,563	3,393,182	3,663,541	3,670,543
Number of non-1kG SNVs	627,319	389,571	515,159	397,852
Number of 1kG indels	212,327	181,597	193,522	195,613
Number of non-1kG indels	145,891	139,035	141,363	221,858
% SNVs that are homozygous	36.11%	39.45%	40.42%	43.99%
Ti/Tv for SNVs in 1kG	2.15	2.15	2.15	2.14
Ti/Tv for SNVs not in 1kG	1.69	1.60	1.65	1.55
Private Protein-Altering (PPA) SNVs	508.6	258	390.9	269
PPA SNVs in COSMIC	3.2	0.5	2.6	1
PPA SNVs in Cancer Genes	13.4	4	8.7	4
PPA indels	24.8	10.5	17.5	35*
PPA indels in COSMIC	0	0	0	0
PPA indels in Cancer Genes	0	0	0.1	1
Total bases in homozygous tracts	49,360,241	52,718,117	83,552,819	374,139,228

**Table S3 | Summary of variants and regions of homozygosity for HeLa and control genomes.**

Variants with a minimum of 8X coverage were annotated as protein-altering using the SeattleSeq annotation server. Private protein-altering (“PPA”) variants were those not observed among the 1000 Genomes Project (“1kG”) or the Exome Sequencing Project 6500 call set, and found outside regions annotated for excessive sequence depth (HiSeq top 5%ile coverage track from the UCSC genome browser). For comparison to COSMIC database, the variant allele was required to match exactly. Comparison to CGP used gene-level overlap. \* HeLa CCL-2 has an increased indel call rate due to higher depth of coverage.

Position	Variant (ref>alt)	dbSNP	Het / Hom	Gene	Amino Acid No.	Class	Alteration	PolyPhen	Grantha m	PhastCons	GERP	COSMIC Match	SangerCGP Somatic	SangerCGP Type
chr1:3347438	C>T	none	het	PRDM16	1096	miss.	ALA,VAL	unknown	64	0.768	4.03	gene	gene	translocation
chr1:228476552	T>G	none	het	OBSCN	3434	miss.	CYS,TRP	unknown	215	0.981	-1.98	exact, COSM210235	none	none
chr4:1808852	G>A	none	het	FGFR3	761	miss.	ASP,ASN	unknown	23	1	4.53	gene	gene	missense, translocation
chr9:139418189	C>T	none	hom	NOTCH1	128	miss.	ARG,HIS	unknown	29	0.996	2.29	gene	gene	translocation, missense, other
chr14:99641176	G>A	none	het	BCL11B	667	miss.	PRO,LEU	unknown	98	0.88	3.95	gene	gene	translocation

Position	Variant (ref>alt)	dbSNP	HGDP	Gene	Amino Acid No.	Class	PhastCons	GERP	COSMIC Match Type	SangerCGP Somatic	SangerCGP Type
chr22:41565536	TTCA>T	none	No	EP300	1401	coding	1	5.55	gene	gene	translocation, frameshift, missense, nonsense, other

**Table S6 | HeLa CCL-2 private protein-altering SNVs/indels overlapping COSMIC or SCGC.**

Variant alleles listed were called in HeLa CCL-2 and found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Also shown for each gene is overlap with the Sanger Cancer Gene Census (SCGC)

Copy Number	Size (kbp)	% of Genome*	HRCN	Size in HRCN (kbp)	% of Genome*
1	5,638	0.22%	1:0	5,638	0.22%
2	691,955	27.21%	2:0	256,419	10.08%
			1:1	435,536	17.13%
3	1,556,135	61.19%	3:0	200,761	7.89%
			2:1	1,355,373	53.30%
4	140,458	5.52%	4:0	31,388	1.23%
			3:1	31,060	1.22%
			2:2	78,010	3.07%
5	42,249	1.66%	5:0	18,081	0.71%
			4:1	15,719	0.62%
			3:2	8,449	0.33%
6	15,702	0.62%	6:0	13,292	0.52%
			5:1	689	0.03%
			4:2	1,721	0.07%
			3:3	0	0.00%
7	3,197	0.13%	7:0	2,225	0.09%
			6:1	0	0.00%
			5:2	971	0.04%
			4:3	0	0.00%

\*Of high-quality, alignable regions

**Table S8 | Haplotype-Resolved Copy Number (HRCN) profile of HeLa CCL-2.**

Proportion of the genome (UCSC hg19/GRC37h, excluding assembly gaps and segmental duplications) at each haplotype-resolved copy number (HRCN) state.

### Chromosome-arm sized LOH regions

Chromosome	Region	Size (Mb)	CN=1?
2q	106,690,345-qter	136.4	No
3q	94,582,003-qter	103.3	No
5p	pter-centromere	46.1	No
6p,6q	pter-qter	170.7	No
11q	102,239,620-qter	32.7	No
13q	19,167,980	95.9	No
19p	pter-12,893,034	12.9	No
22q	16,385,650-qter	34.8	No
Xp,Xq	pter-qter	152.1	No

### Short LOH regions

Chromosome	Region	Size (kb)	CN=1?
2	40,339,750-41,992,745	1,653	No
3	80,281,400-81,385,274	1,104	Yes
4	158,267,826-161,280,735	3,013	Yes
4	172,475,207-173,703,673	1,228	No
7	15,684,943-16,895,503	1,211	No
7	123,832,242-126,478,678	2,646	No
11	184,961-2,876,557	2,692	Yes
11	22,372,309-24,503,054	2,131	No

**Table S9 | Large regions of LOH in HeLa CCL-2.**



IN 1000 GENOMES?	Yes	No
Allele not observed (depth $\geq 2$ ) in clones	179000	107501
Allele observed only in unphased clones	3342	23603
Unphased due to inconsistency (observed in A and B clones with equal scores)	3732	8510
Phased by majority rule among clones, with conflicting phase calls between clones	30496	32326
Phased unanimously among clones, only one allele observed	613890	69806
Phased unanimously among clones, both alleles observed	1143908	62709

IN 1000 GENOMES? IN SEGDUPE? REPEAT-MASKED?	Yes No No	Yes No Yes	Yes Yes No	Yes Yes Yes	No No No	No No Yes	No Yes No	No Yes Yes
Allele not observed (depth $\geq 2$ ) in clones	67404	100728	4997	5871	9051	47410	22787	28253
Allele observed only in unphased clones	827	1637	479	399	6612	10130	3647	3214
Unphased due to inconsistency (observed in A and B clones with equal scores)	1147	1948	315	322	1808	3559	1501	1642
Phased by majority rule among clones, with conflicting phase calls between clones	12708	14986	1394	1408	8071	12333	5540	6382
Phased unanimously among clones, only one allele observed	268193	323668	10230	11799	14376	30783	11705	12942
Phased unanimously among clones, both alleles observed	545114	563384	17552	17858	21966	29179	5541	6023

**Table S12 | Phasing status of heterozygous SNVs in HeLa CCL-2.**

Counts of heterozygous SNVs are shown by phasing status (phased or unphased, and reason) and overlap with 1000 Genomes Project data and genomic repeats (segmental duplications or regions identified by Repeat Masker). For unphased variants, the reason for lack of phase assignment is indicated (does not appear among clones, or alleles are inconsistent among phased clones). Phased variants are separated by the degree of support among clone data (both alleles observed with no inconsistency between clones, or only one allele observed with no inconsistency between clones, or inconsistencies between clones resolved by majority rule).

HRCN (Total : HapA : HapB)	Total genomic extent (bp)	Total bp of duplicated haplotype(s) (extent x copy)	Number clone-confirmed mutations	Clone-confirmed somatic mutation frequency (per bp x 10 <sup>6</sup> )	Expected frequency given 61% sensitivity (per bp x 10 <sup>6</sup> )
2:2:0	369,962,202	739,924,404	1022	1.38	2.26
3:3:0	328,232,258	984,696,774	1437	1.46	2.39
4:4:0	54,229,430	216,917,720	287	1.32	2.17
5:5:0	11,618,893	58,094,465	39	0.67	1.10
6:6:0	33,636,597	201,819,582	98	0.49	0.79
3:2:1	1,395,662,889	2,791,325,778	4128	1.48	2.42
4:2:2*	221,271,991	885,087,964	891	1.01	1.65
4:3:1	64,697,997	194,093,991	216	1.11	1.82
5:3:2*	2,368,480	11,842,400	7	0.59	0.97
5:4:1	19,813,202	79,252,808	30	0.38	0.62
6:4:2*	5,861,548	35,169,288	10	0.28	0.47
<b>TOTAL</b>	<b>2,507,355,487</b>	<b>6,198,225,174</b>	<b>8165</b>	<b>1.32</b>	<b>2.16</b>

**Table S14 | Clone-confirmed somatic mutation frequency.**

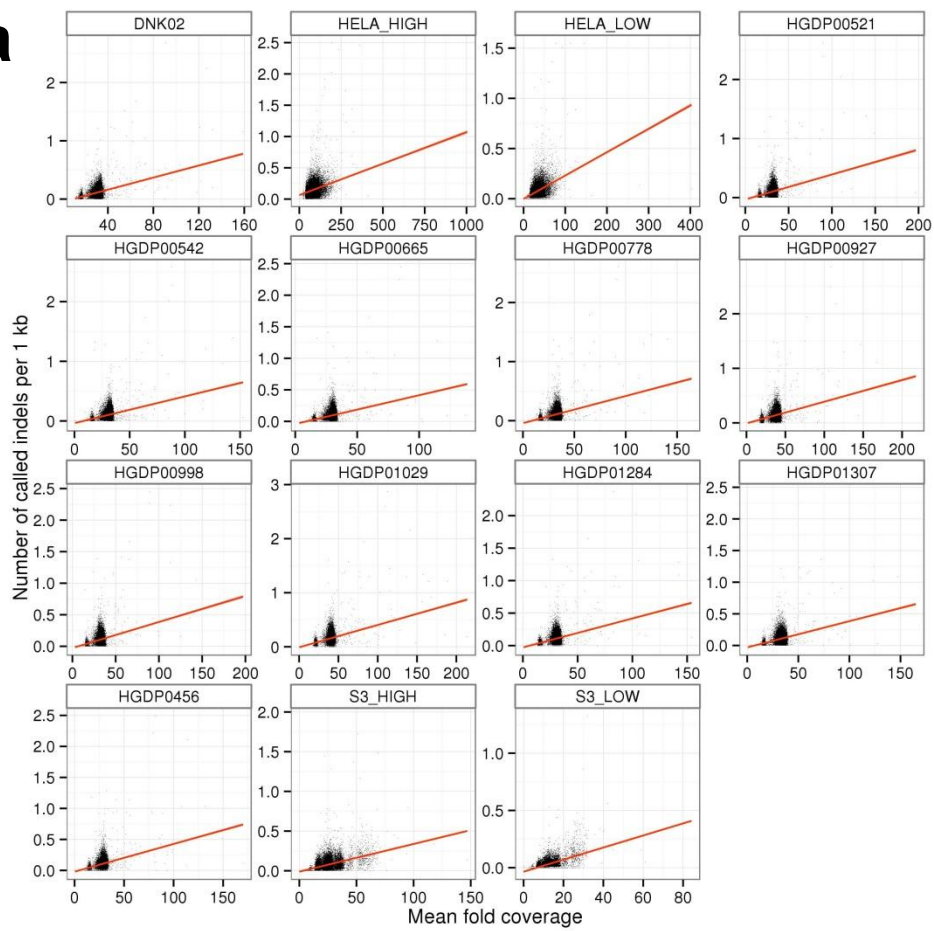
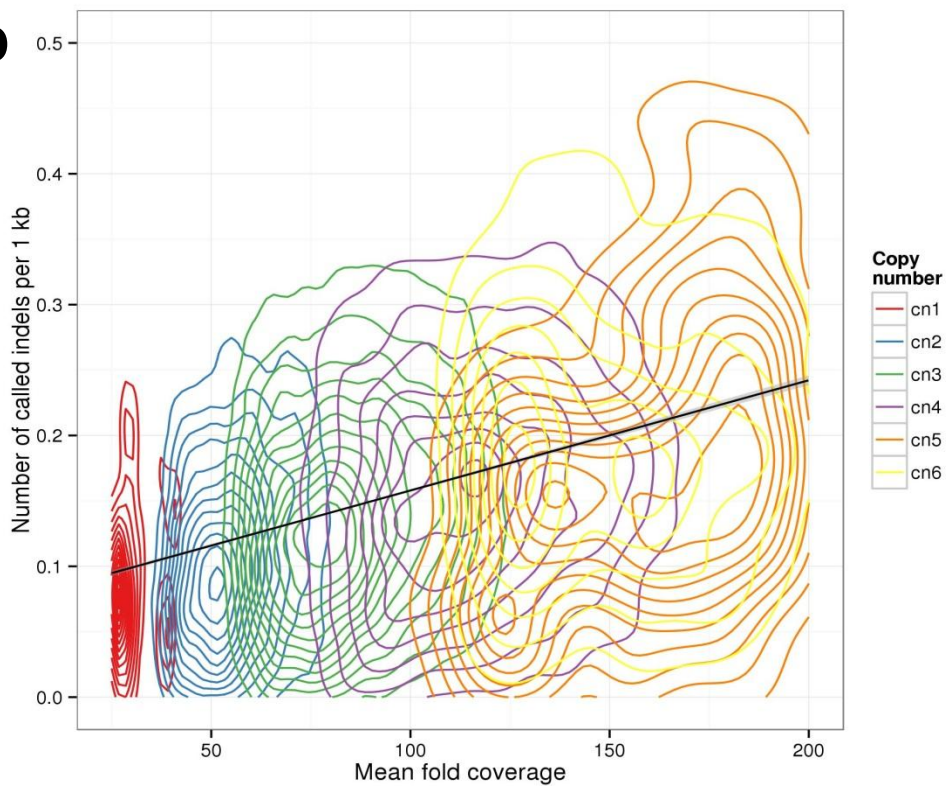
Counts and frequencies of somatic mutations in the HeLa CCL-2 genome. The total number of bases in the genome at each haplotype-resolved copy number (HRCN) state (total copies:haplotype A copies:haplotype B copies) are listed, as well as the number of somatic mutations observed and confirmed by clone pool sequencing. Mutations occurring on duplicated haplotypes could arise on any of the haplotype copies, so mutation rate is taken as (# sites in given C.N.) / ( [total bases within reference at C.N.] x [copies of duplicated haplotype(s)]). Shaded rows indicate regions of LOH (haplotype B copies = 0).

\*In these regions, both haplotypes are duplicated, so mutations on either were considered; in all other cases, only mutations occurring on the major haplotype were counted.

ID	Genotyped for:	Num. $\geq 8X$ (both)	Num. Shared	Percent Shared
S3 DNA	CCL-2 SNVs	204,800	194,416	94.93
S3 DNA	CCL-2 protein-altering SNVs	301	249	82.72
S3 RNA	CCL-2 SNVs	22,772	22,129	97.12
S3 RNA	Shared S3 & CCL-2 protein-altering SNVs	74	65	87.84
CCL-2	S3 SNVs	55,540	50,610	91.12
CCL-13	CCL-2 SNVs	47,781	43,507	91.06
CCL-5	CCL-2 SNVs	55,696	50,596	90.84
CCL-17	CCL-2 SNVs	45,734	41,847	91.50
CCL-23	CCL-2 SNVs	41,668	37,914	90.99
CCL-25	CCL-2 SNVs	44,262	40,632	91.80
CCL-6	CCL-2 SNVs	37,119	33,623	90.58
CCL-62	CCL-2 SNVs	42,249	38,481	91.08
CCL-21	CCL-2 SNVs	38,906	35,476	91.18

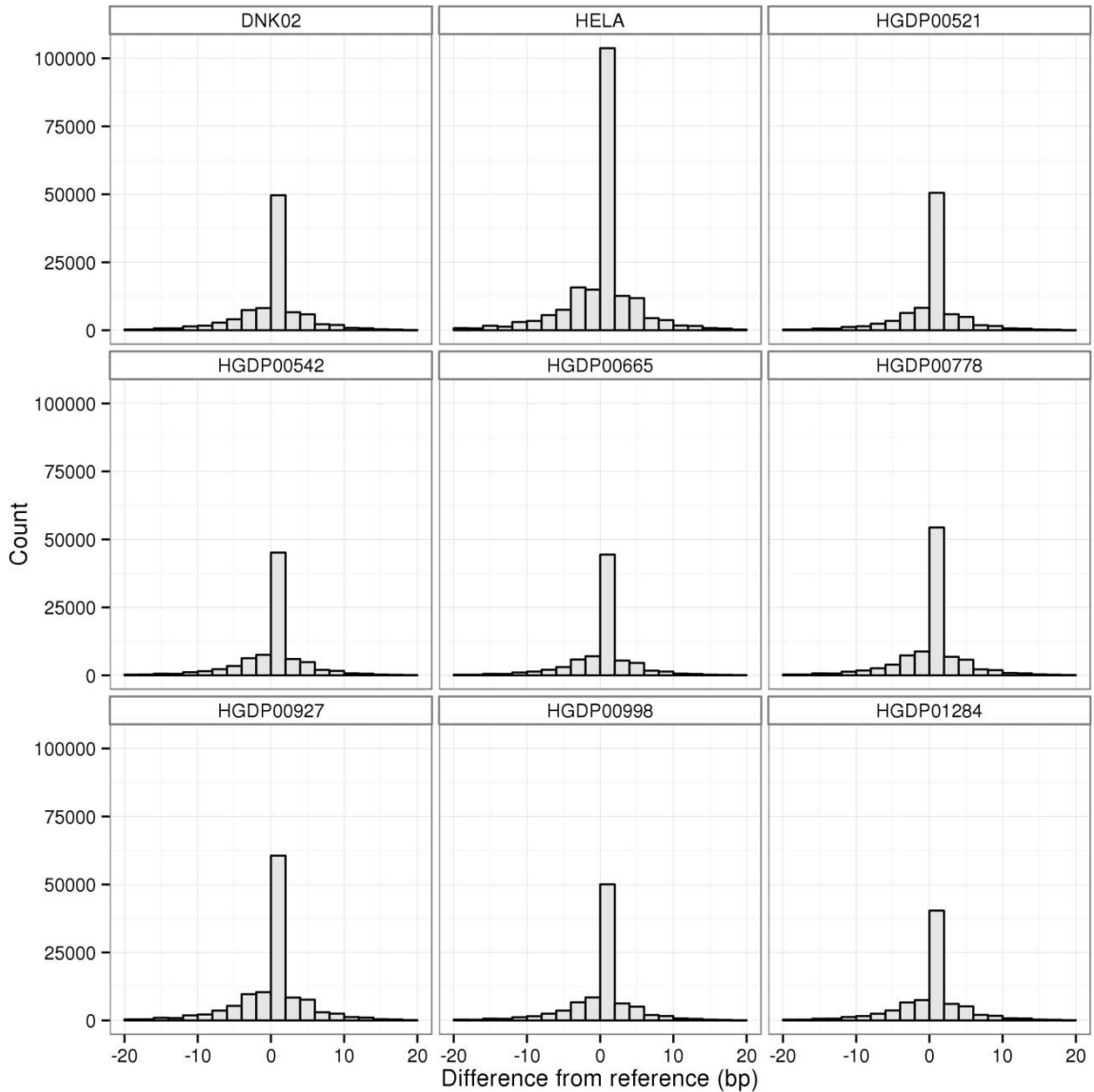
**Table S15 | Variants shared between HeLa strains**

HeLa S3 shotgun reads, HeLa S3 RNA-Seq reads, and shotgun reads from 8 additional HeLa strains were genotyped at HeLa CCL-2 variant sites for the presence or absence of the HeLa CCL-2 variant allele. Positions were only included if both HeLa CCL-2 and the data set have a coverage of at least 8x, and are not in segmental duplications or at 1000 Genomes Project sites.

**a****b**

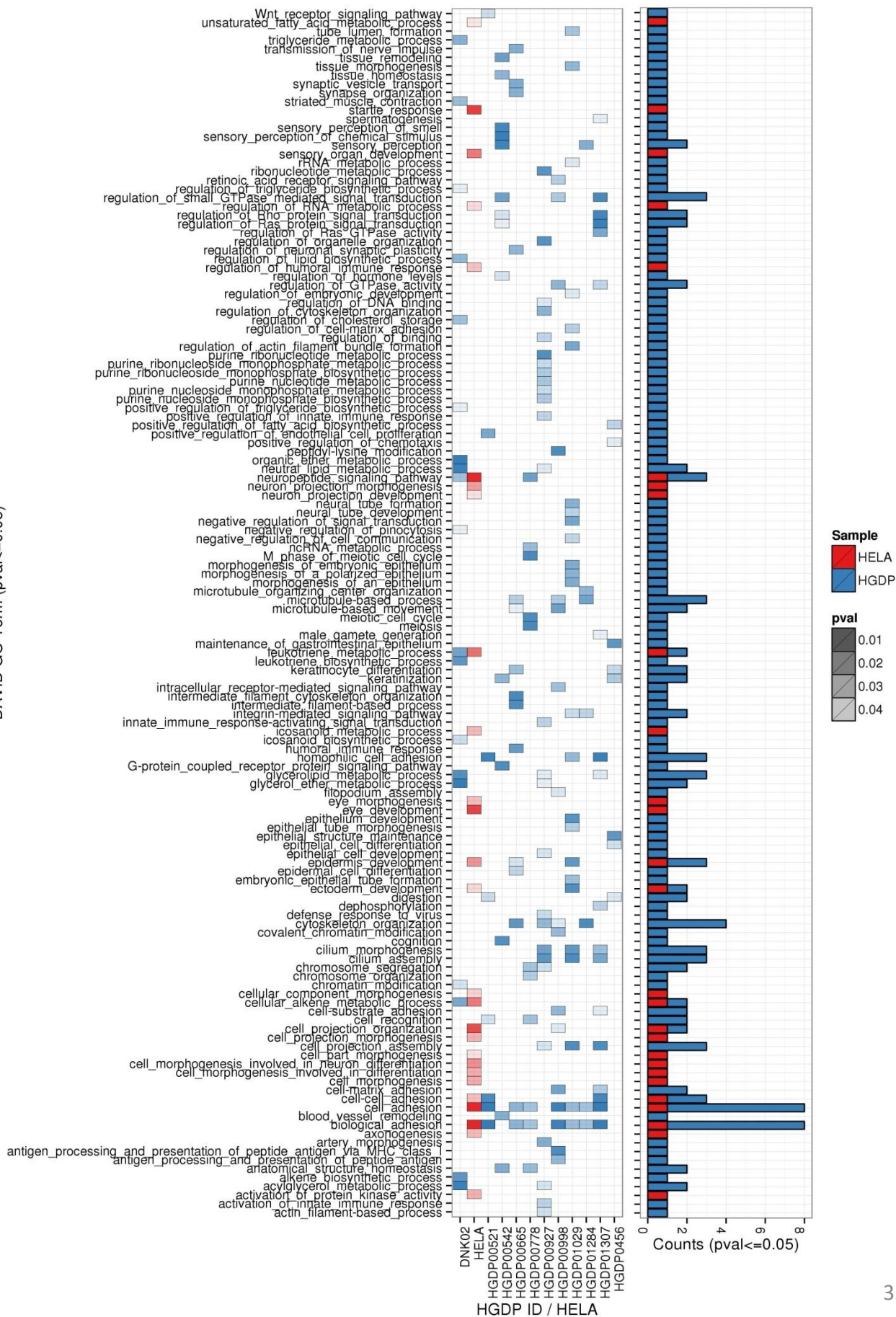
## Figure S1 | Indel calling by coverage

**a.** Counts of indels called is plotted versus read depth at each indel for HeLa and 11 HGDP controls. Shotgun reads from HeLa CCL-2 as well as HeLa S3 were randomly downsampled in order to study the effects of lower coverage upon indel counts in each genome. Mean coverage was HeLa CCL-2 full dataset: ("HELA\_HIGH"), ~88X; HeLa CCL-2, subsampled ("HELA\_LOW"), ~35X; HeLa S3 full dataset ("S3\_HIGH"), ~26X; HeLa S3 subsampled ("S3\_LOW"), ~12X; 11 HGDP controls, ~30-45X. Each point represents one of the low resolution SUNK windows (mean size, 77 kbp), and for each window, mean read depth and total number of indel calls per kilobase were determined. In all genomes analyzed, there is a strong correlation between number of calls by read depth. **b.** Indel calls in HeLa (88X) for points as in **a** but shown as a 2d density contour plot, split by underlying copy number. As the mean coverage increases with the copy number so does the ability to call indels, resulting in a higher call count per kilobase at higher copy numbers.



**Figure S2 | STR profiling with lobSTR.**

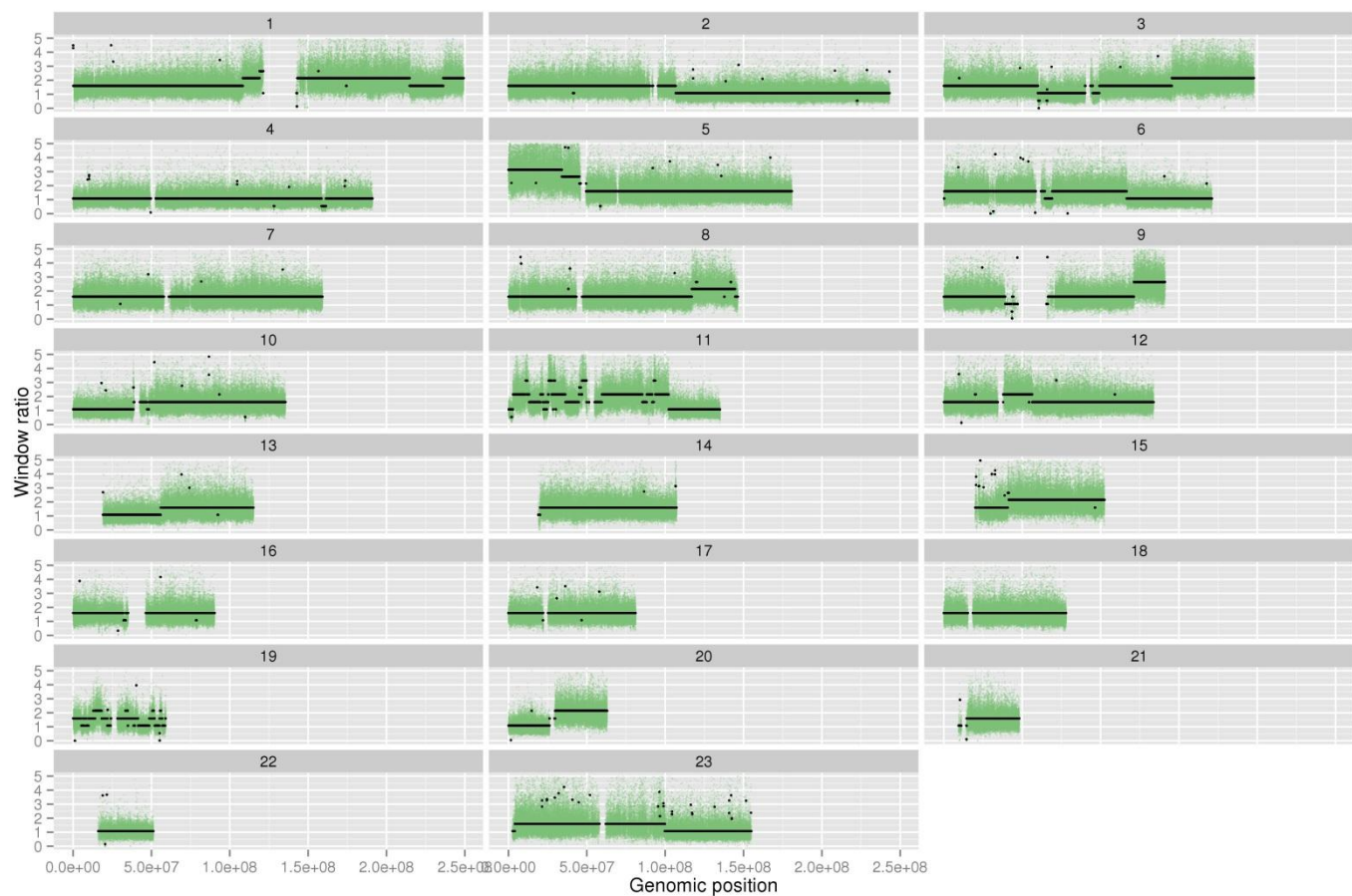
Short Tandem Repeats (STRs) were identified using lobSTR (Gymrek, M, *et. al.* (2012)) for HeLa as well as eight of the diversity panel control individuals (Rohland, N; Reich, D (2012)). Repeats with a coverage of at least 10 are represented above as a histogram of counts for the length difference in base pairs of called STRs from the reference. While more calls above the coverage threshold are called for HeLa, likely due to having 88X coverage compared to ~30-45X for the control samples, the profile of lengths are comparable between all samples.



**Figure S3 | Gene ontology enrichment analysis for genes with protein-altering variants in HeLa CCL-2 and 11 HGDP controls.**

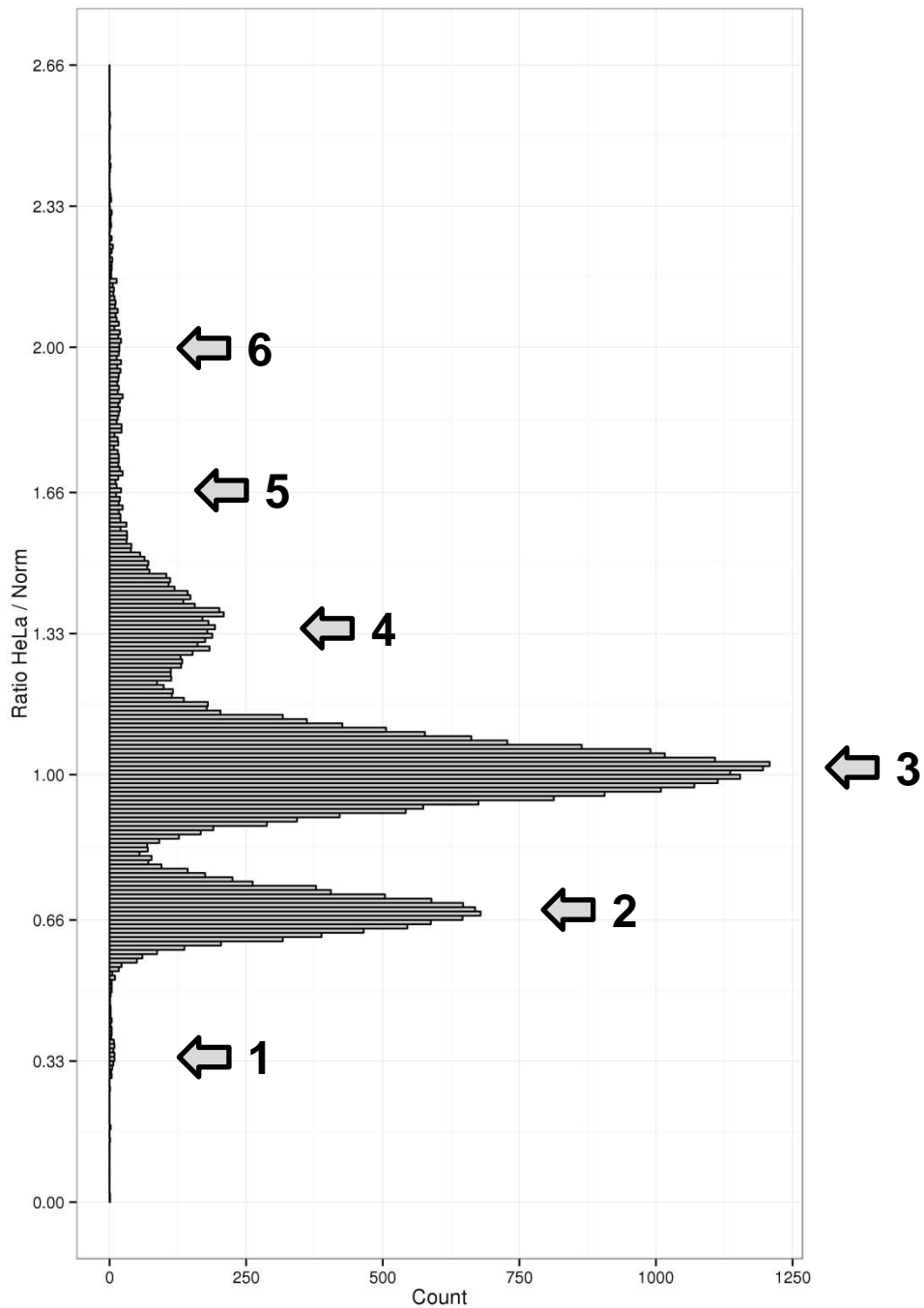
For HeLa CCL-2 and the 11 control genomes, a list of genes with protein-altering SNVs, indels, structural rearrangements, or copy number alterations (copy-number <1 or >9) was analyzed by DAVID (Huang *et. al.* (2009)). Gene Ontology terms (GO-terms) were then filtered to retain only those with a p-value  $\leq 0.05$  and plotted in the left panel where color indicates the genome (HeLa or control) and shading represents significance. The right panel shows, for each term, the number of genomes with significant enrichment for protein-altering variants in the associated genes. With the exception of the “Startle response” GO-term, all of the terms in HeLa with a p-value  $\leq 0.01$  occur in at least one of the control genomes.





**Figure S4 | HeLa CCL-2 high resolution copy number calls.**


Copy number ratios versus control genomes are plotted within high-resolution SUNK windows (green dots, each window size ~1.5 kb), with predicted copy number state overlaid (black dots).



**Figure S5 | HeLa over GC-matched control ratio histogram.**

SUNK window (500 unique 30mer) resolution ratio scores plotted as a histogram. Distinct peaks are observed at approximately 0.33, 0.66, 1.0, 1.33, consistent with an approximately triploid numerator sample (HeLa) over a diploid denominator sample (GC-matched control). Inferred copy numbers are indicated by arrows.

**a**

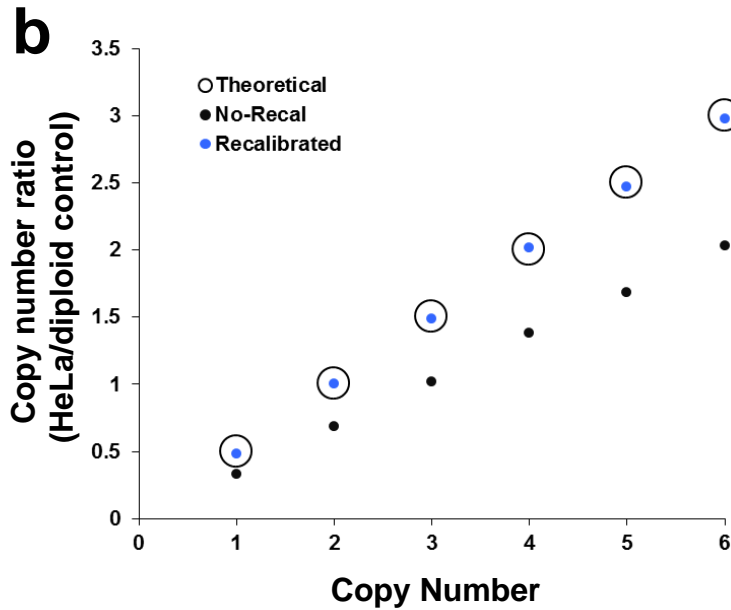
Actual, unknown copy number																					
Aneuploid	2	2	2	2	4	4	4	2	2	1	2	2	6	6	4	4	4	2	2	2	= 59
"Normal"	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	= 40
Total amount of "genetic material" 																					

Resulting Normalized Ratio Scores																				
0.68	0.68	0.68	0.68	1.36	1.36	1.36	0.68	0.68	0.34	0.68	0.68	2.04	2.04	1.36	1.36	1.36	0.68	0.68	0.68	= 20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	= 20
Normalization Constant →																				

HMM State Segmentation									
HMM State ID		1	=	1	=	0.34	Resulting state mean ratio values. Copy numbers are still unknown.		
		2	=	2	=	0.68			
		3	=	4	=	1.36			
		4	=	6	=	2.04			
		Mean state score							
		Copy number (still unknown)							

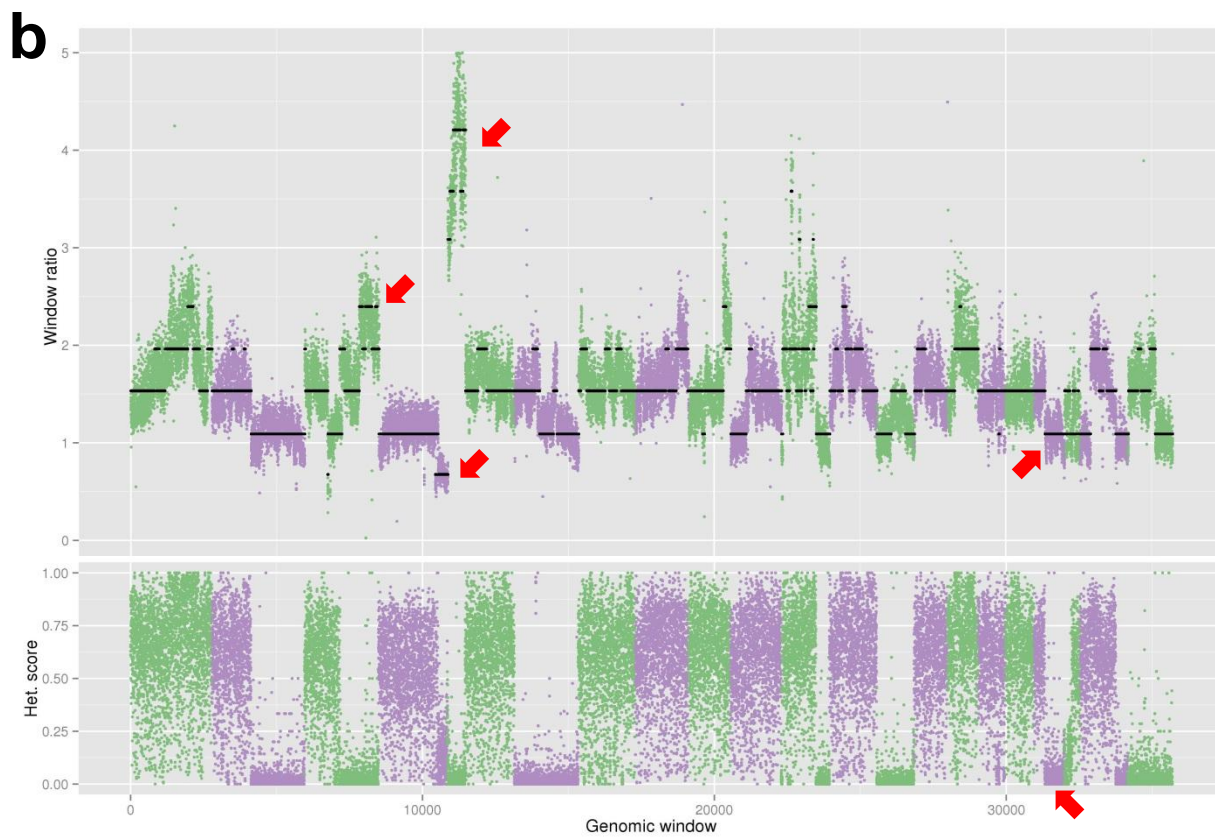
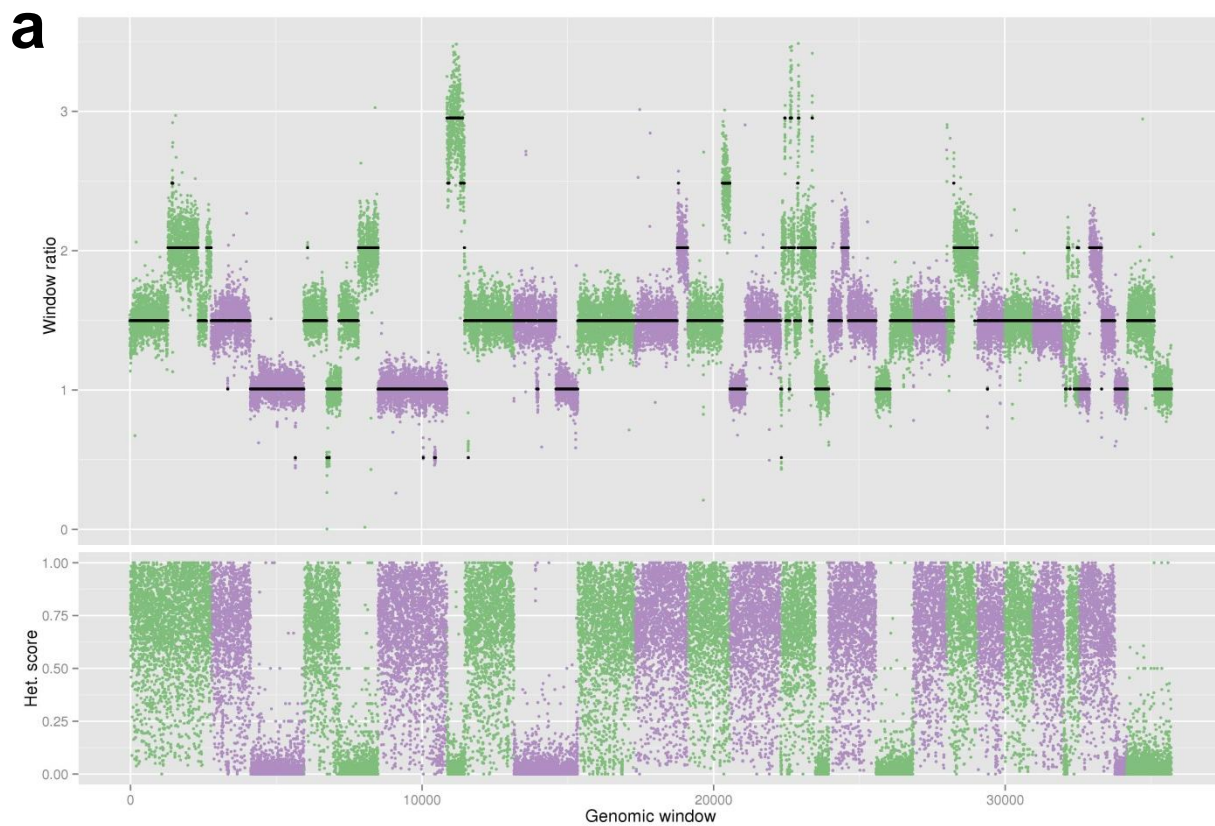
Hypotheses generated & copy numbers assigned																						
Incorrect Hypotheses	2	2	2	2	3	3	3	2	2	1	2	2	4	4	3	3	3	2	2	2	= 49	= 1.225
	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	= 40	
Correct Hypotheses	2	2	2	2	4	4	4	2	2	1	2	2	6	6	4	4	4	2	2	2	= 59	= 1.475
	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	= 40	

New state means calculated and compared to theoretical																			
Incorrect Hypotheses		1	=	1	=	0.6125	!=	0.5											
		2	=	2	=	1.125	!=	1.0											
		3	=	3	=	1.875	!=	1.5											
		4	=	4	=	2.45	!=	2.0											
Correct Hypotheses		1	=	1	=	0.5	==	0.5											
		2	=	2	=	1.0	==	1.0											
		3	=	4	=	2.0	==	2.0											
		4	=	6	=	3.0	==	3.0											



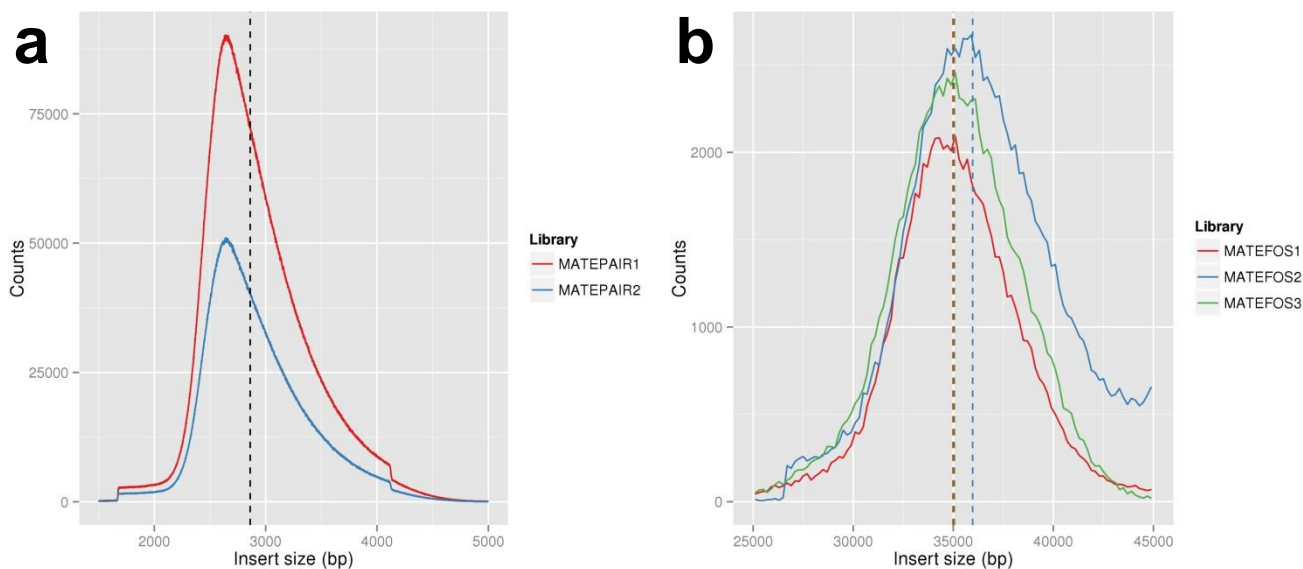
**Figure S6 | Copy-number recalibration strategy.**

**a.** Schematic of steps involved in the recalibration process. In order to adjust for differences in total read depth between genomes, window scores are normalized to a constant. Ratios are then taken between a G+C profile matched normal control and states are segmented using an HMM. Resulting ratios are not directly relatable to absolute copy number when the two genomes' chromosomal complements are of unequal size (e.g., one is triploid and the other diploid). Assignments of copy numbers to HMM states ("hypotheses") are exhaustively generated; windowed copy number values then summed to generate a "genetic material ratio" which is used as the normalization constant. The mean across windows from each HMM state is recalculated, and ratios to the diploid control genome are taken, after which the per-state . The hypothesis which minimizes the mean difference between observed and expected per-state ratios is chosen. **b.** HeLa copy number state scores are shown before and after recalibration (black and blue, respectively), with theoretical values shown as open circles.



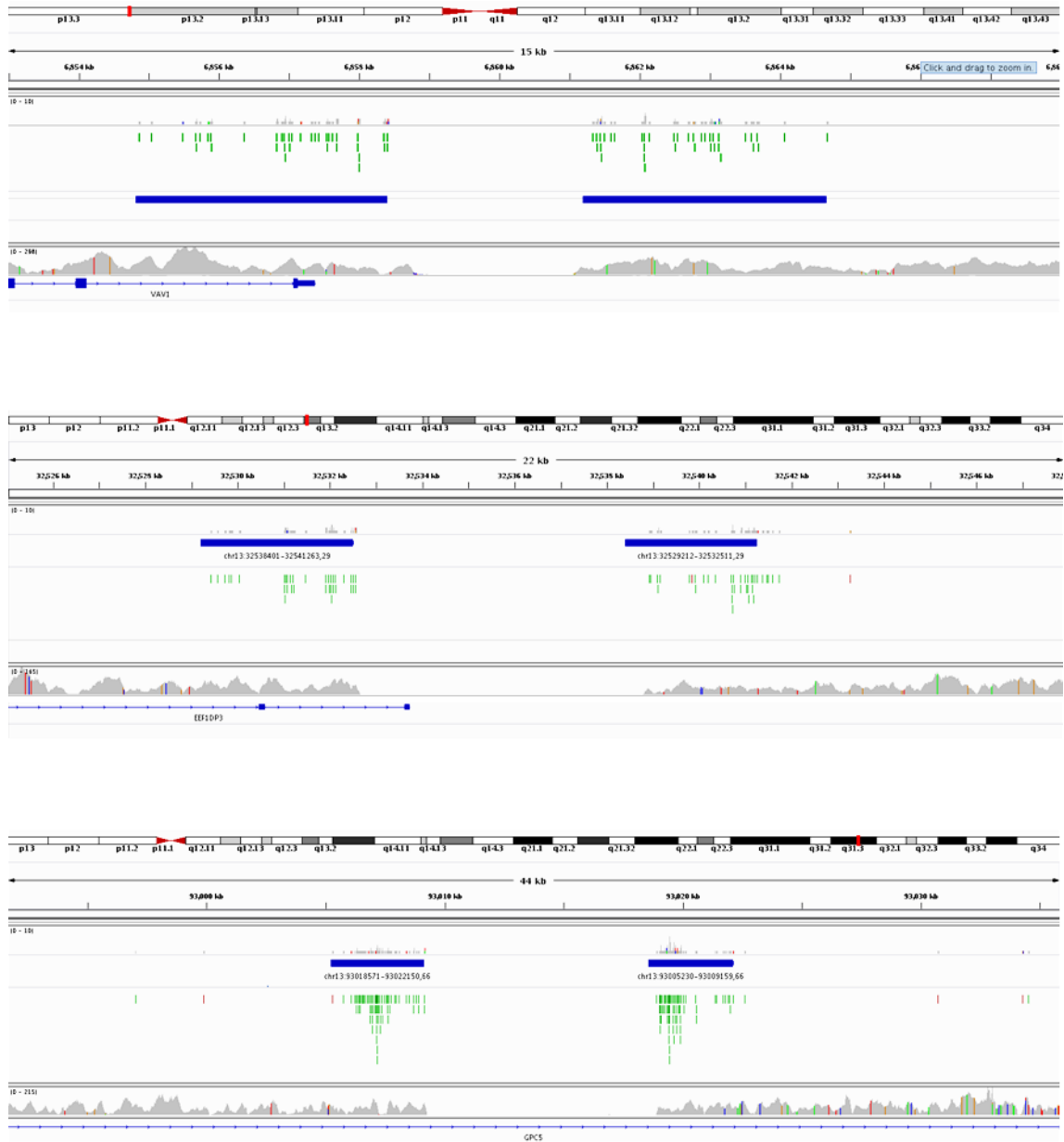
**Figure S7 | HeLa CCL-2 and S3 copy number and LOH profiles.**

**a**, Top - Low resolution SUNK window ratio scores (green or purple points) and copy number state calls (black lines) for HeLa CCL-2. Bottom – Loss of heterozygosity scores measured by the fraction of heterozygous variants in each window. **b**, As in **a**. but for HeLa S3. Red arrows indicate notable changes in copy number or LoH.



**Figure S8 | Mate pair insert size distributions.**

**a**, Insert size distributions of concordant pairs for the two "3 kb" mate-pair libraries constructed using *in vitro* circularization (Talkowski *et. al.* (2011)). **b**, Insert size distributions of concordant pairs for the three "40 kb" mate-pair libraries constructed using fosmid cloning.



**Figure S9 | Examples of deletions in HeLa CCL-2**

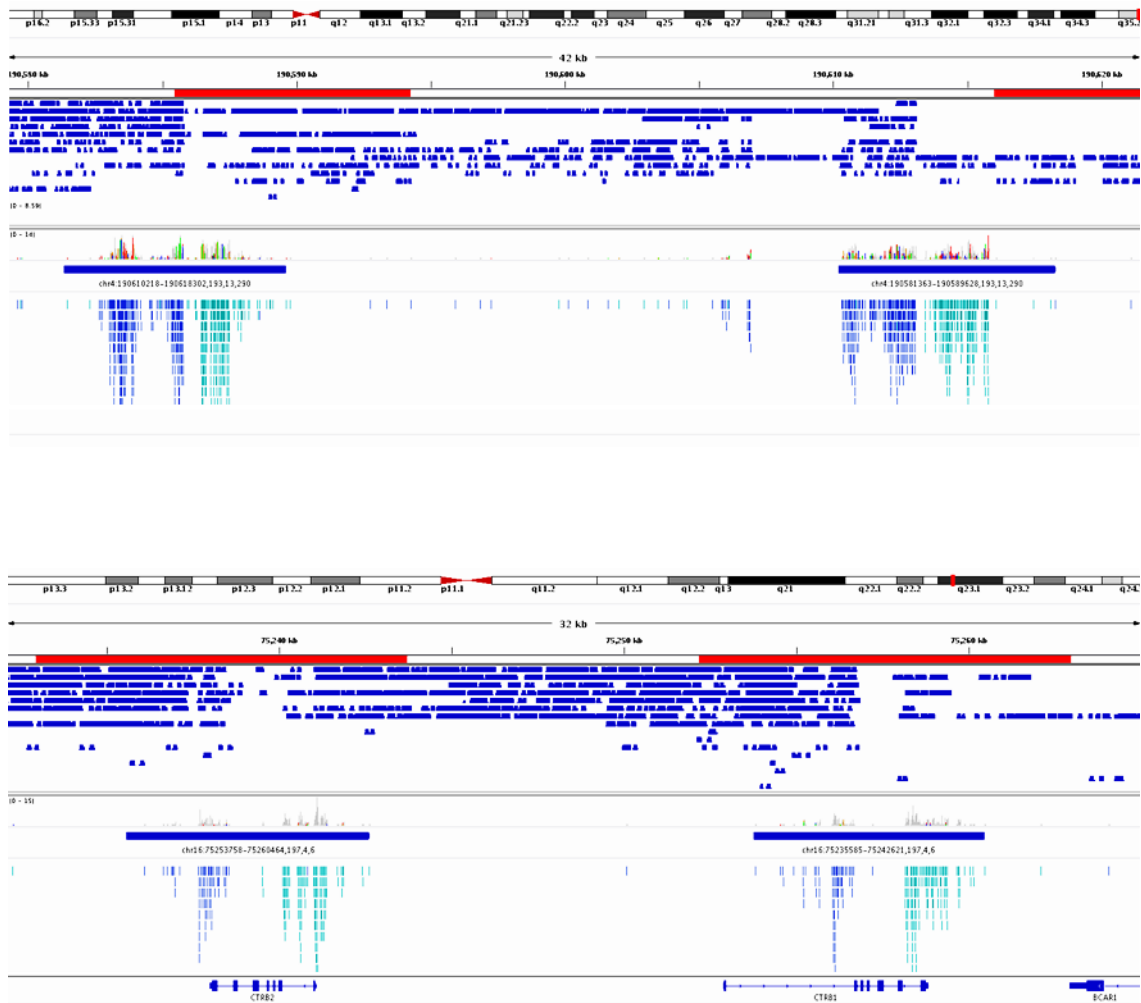
Three examples of deletions called using a sliding window approach shown in the IGV genome browser. Blue bars denote regions of coverage from supporting 3 kb mate-paired reads (green ticks). Shotgun sequence coverage (gray bars) are plotted beneath each event.





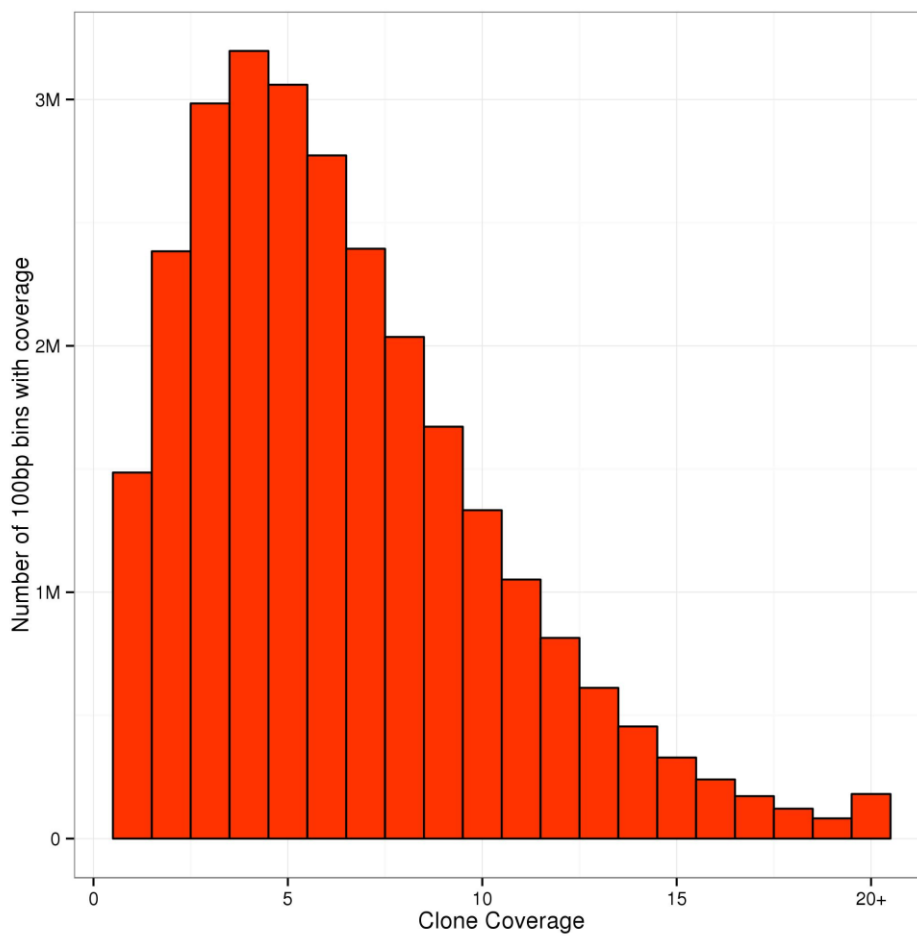
**Figure S10 | Examples of inter-chromosomal rearrangements in HeLa CCL-2.**

Two examples of inter-chromosomal rearrangements detected by a sliding window approach from discordantly-mapping 3 kb mate-pair reads. The upper example is one of the rearrangements within marker chromosome M14.



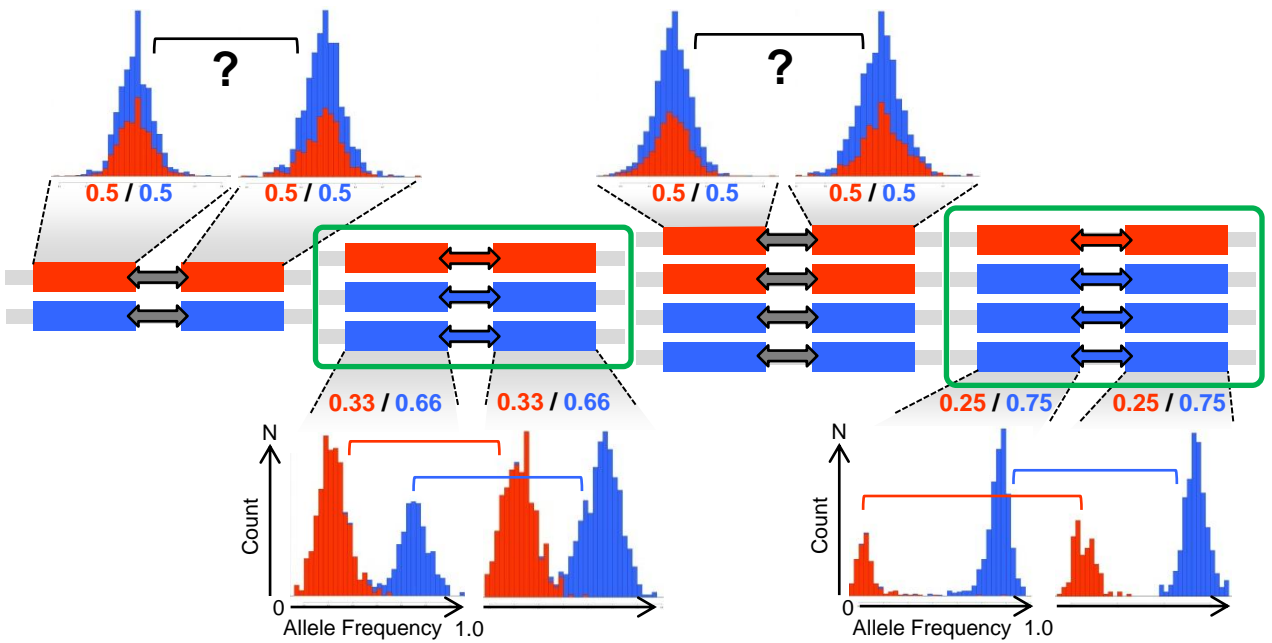
**Figure S11 | Called inversion examples in HeLa CCL-2.**

Two examples of inversions detected by a sliding window approach from discordantly-mapping 3 kb mate-pair reads. Both inversions are supported by fosmid sequence coverage profiles (blue tracks shown below chromosome ideograms), with overlapping clones showing discontinuous patterns of coverage near each inversion breakpoint.



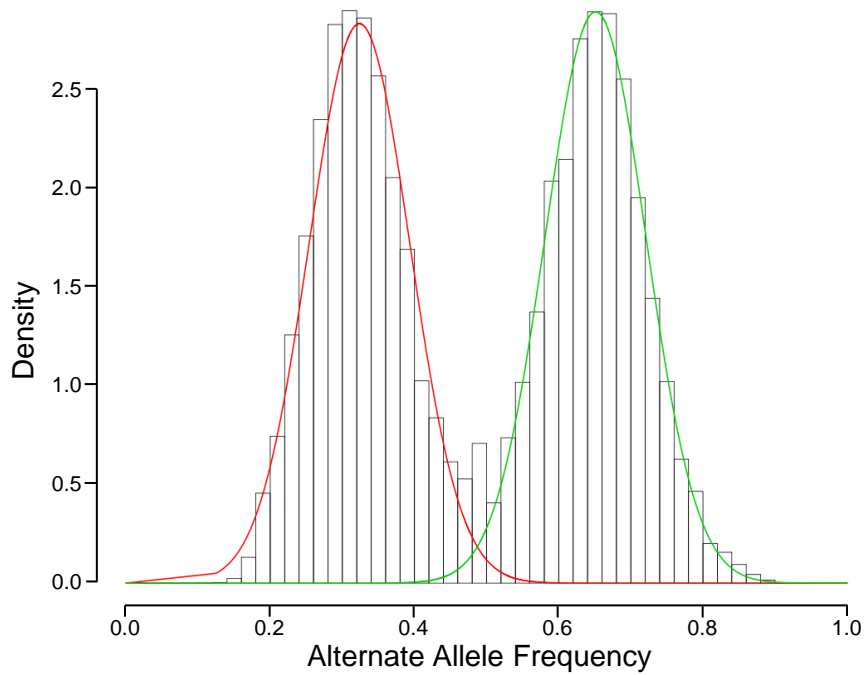
**Figure S12 | Histogram of clone coverage.**

Histogram of the physical coverage by fosmid clone inserts. Overall, 3.5% of the genome is not covered (coverage=0, excluding chromosome Y and assembly gaps).



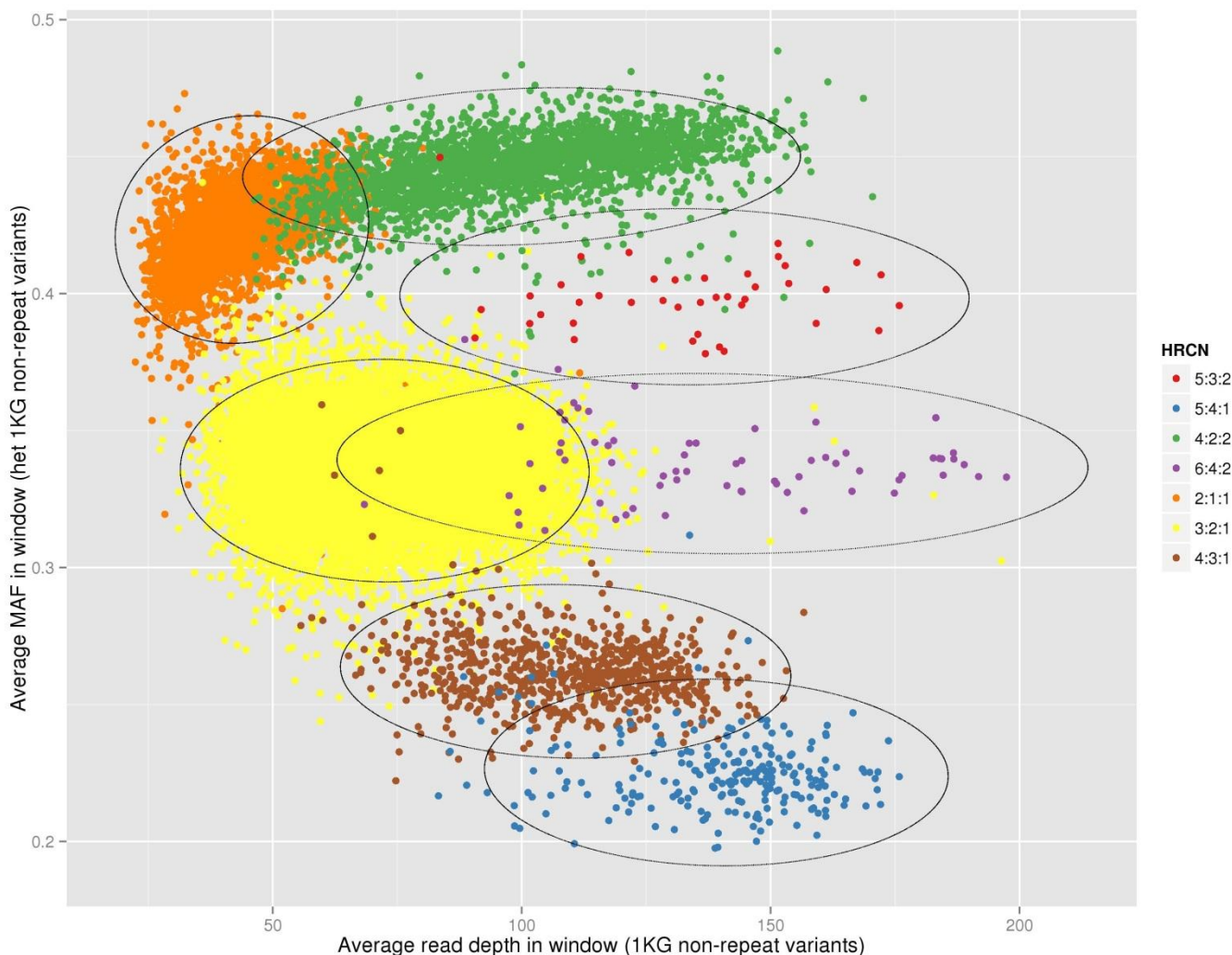
**Figure S13 | Schematic of haplotype scaffolding approach using allele imbalance.**

Consecutive haplotype blocks in regions containing imbalanced copy numbers (green boxes) between haplotypes can be merged using an HMM to form a haplotype scaffold based on the allele frequencies of phased variants within the blocks (histograms of with (haplotype A) and blue (haplotype B) distributions representing allele frequencies for the respective haplotypes). For haplotype blocks in regions of imbalanced haplotype these histograms are distinct (histograms on bottom of figure), whereas haplotype blocks in regions of even copy number overlap and can not be distinguished (histograms at the top of the figure).



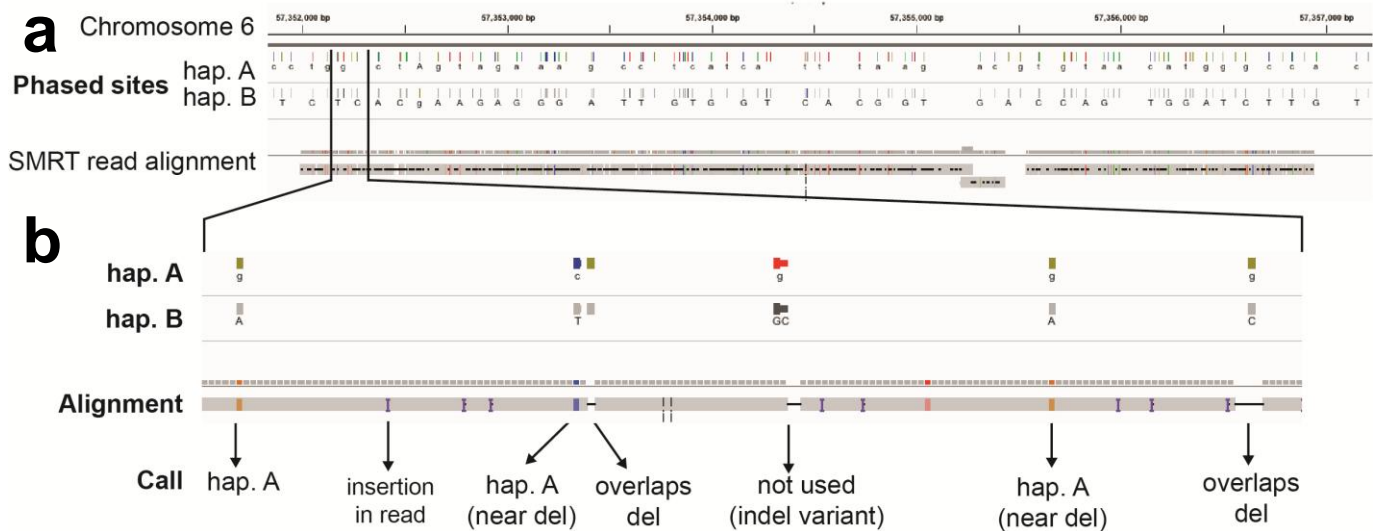
**Figure S14 | Gaussian mixture model of AAFs in non-LOH copy number 3 regions.**

A histogram of alternate allele frequencies among shotgun reads is shown for all heterozygous variants present in regions of copy number 3 in which one haplotype is at copy number 2 and the other at copy number 1. A two-component Gaussian mixture model was fit to this distribution, and the centers of each component (red and green lines) were at 0.324 and 0.651, near the expected values of  $1/3$  and  $2/3$ .



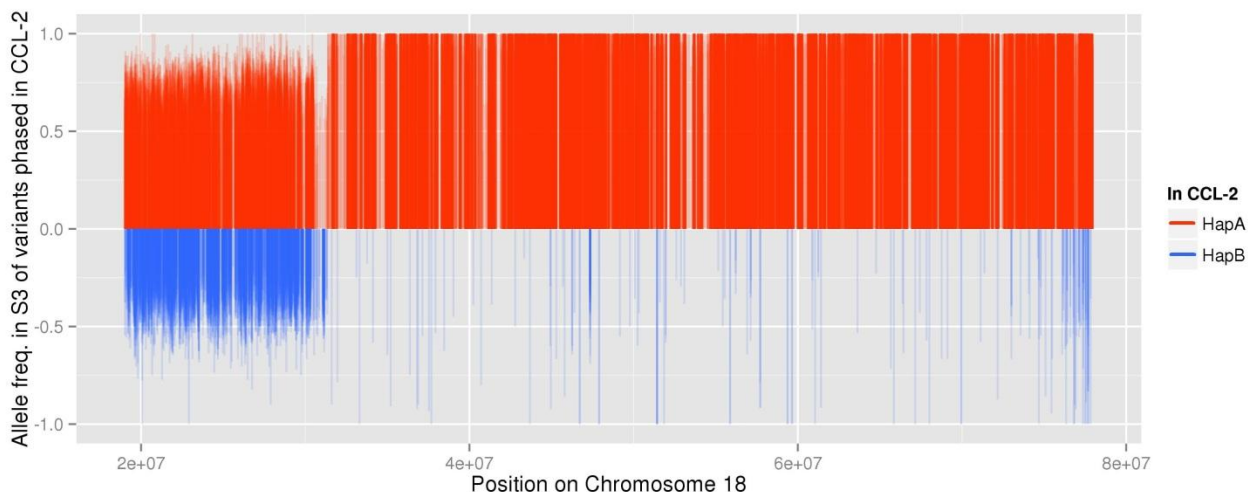
**Figure S15 | HeLa allele balance by read depth for HRCN regions.**

For each low resolution SUNK window (~77 kb), the average minor allele frequency of all heterozygous variants was plotted against those sites' average read depth. Each point was shaded by the window's predicted HRCN (total copy number : haplotype A copy number : haplotype B copy number). Overlaid ellipses represent 95% confidence intervals for each HRCN grouping.



**Figure S16 | Long read haplotype validation.**

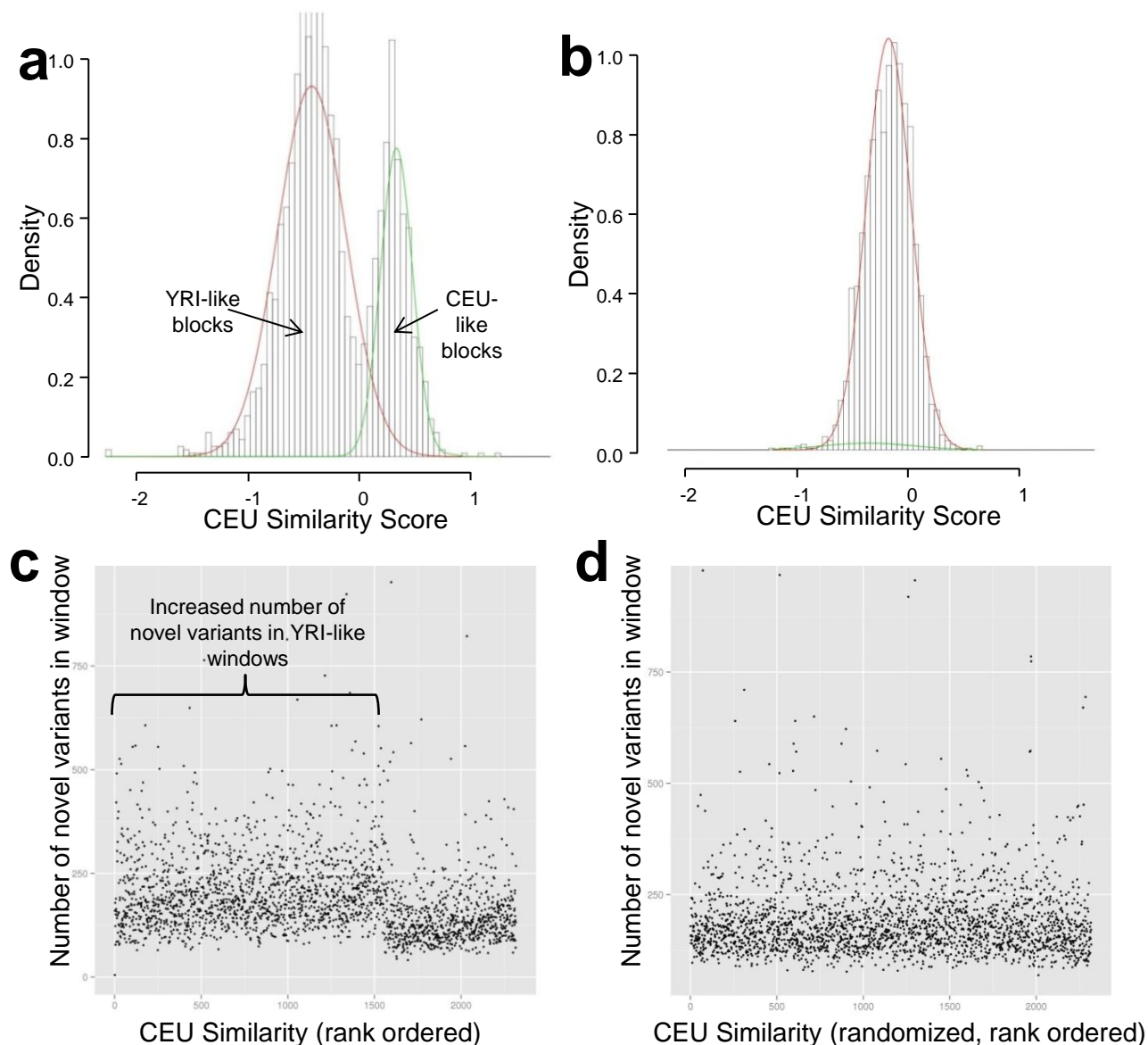
Haplotype phase validation by single molecule long-read sequencing. An example alignment from one read spanning 4.96 kbp is shown. **a.** Upper track: phased variants in HeLa CCL-2 are shown for each inherited haplotype (A and B), with gray ticks indicating the reference allele and colors representing the alternate allele. Lower track: the aligned read spans 98 phased heterozygous sites, of which 19 sites are more than 10 bp from the nearest alignment indel. Of those, all 19 sites match the allele predicted on haplotype A. **b.** Detail showing aligned positions matching haplotype A or rejected due to overlapping or nearby indel errors.



**Figure S17 | Allelic state across LOH event specific to HeLa S3.**

Allele frequencies among HeLa S3 shotgun reads are shown for all heterozygous and phased variants from HeLa CCL-2, across 78.1 Mbp of chromosome 18. Allele frequencies in S3 are plotted on the y-axis, with points' direction and color indicating whether each CCL-2 allele is phased on haplotype A (red, upward) or haplotype B (blue, downward). In HeLa CCL-2, chromosome 18 is triploid without LOH, but in HeLa S3 it is observed to have a large (47.3 Mbp) distal region that is diploid with LOH. Nearly all (99.7%) of the variants with allele balance  $>0.9$  within this region (in S3) correspond to haplotype A from HeLa CCL-2.

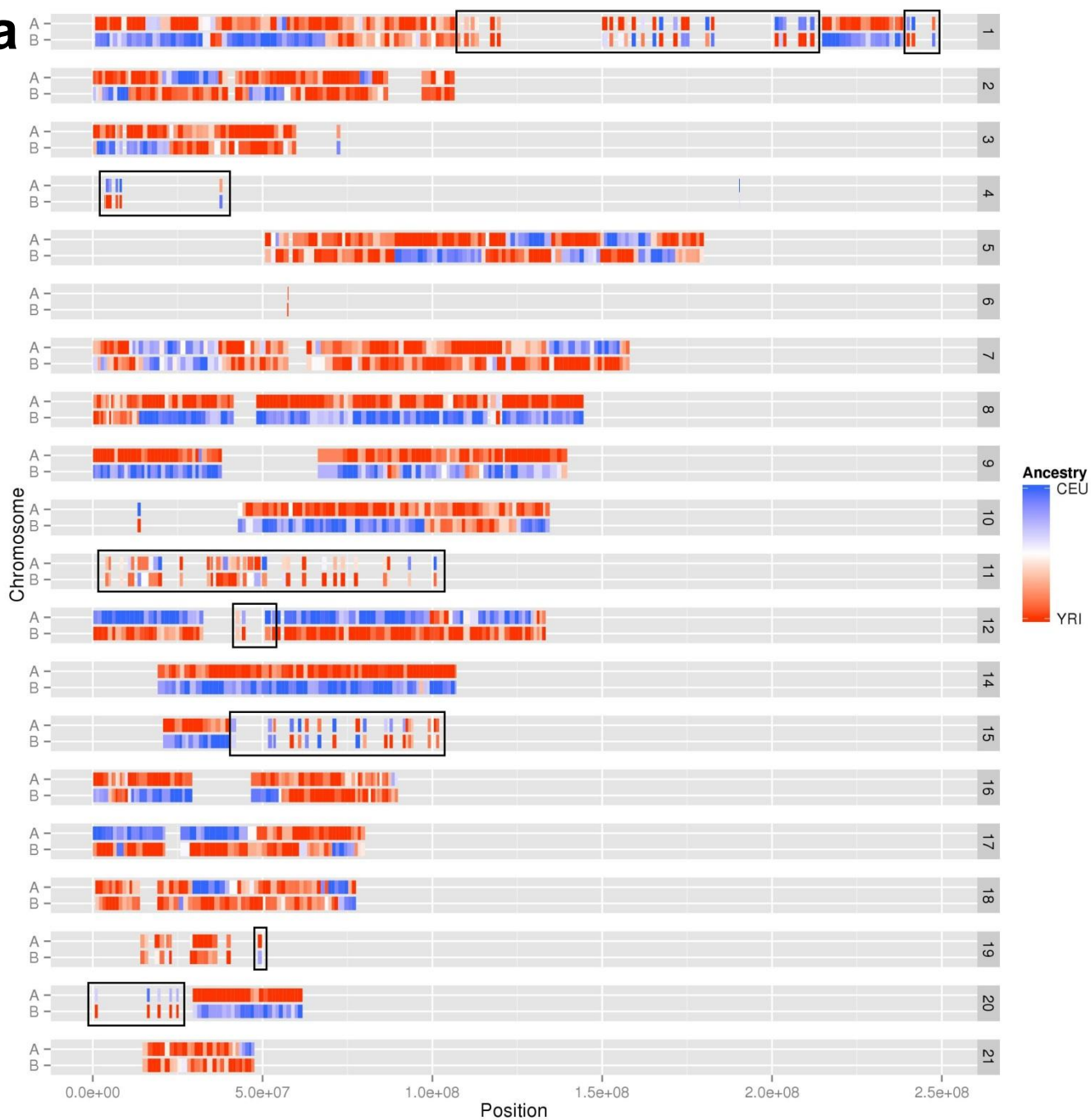


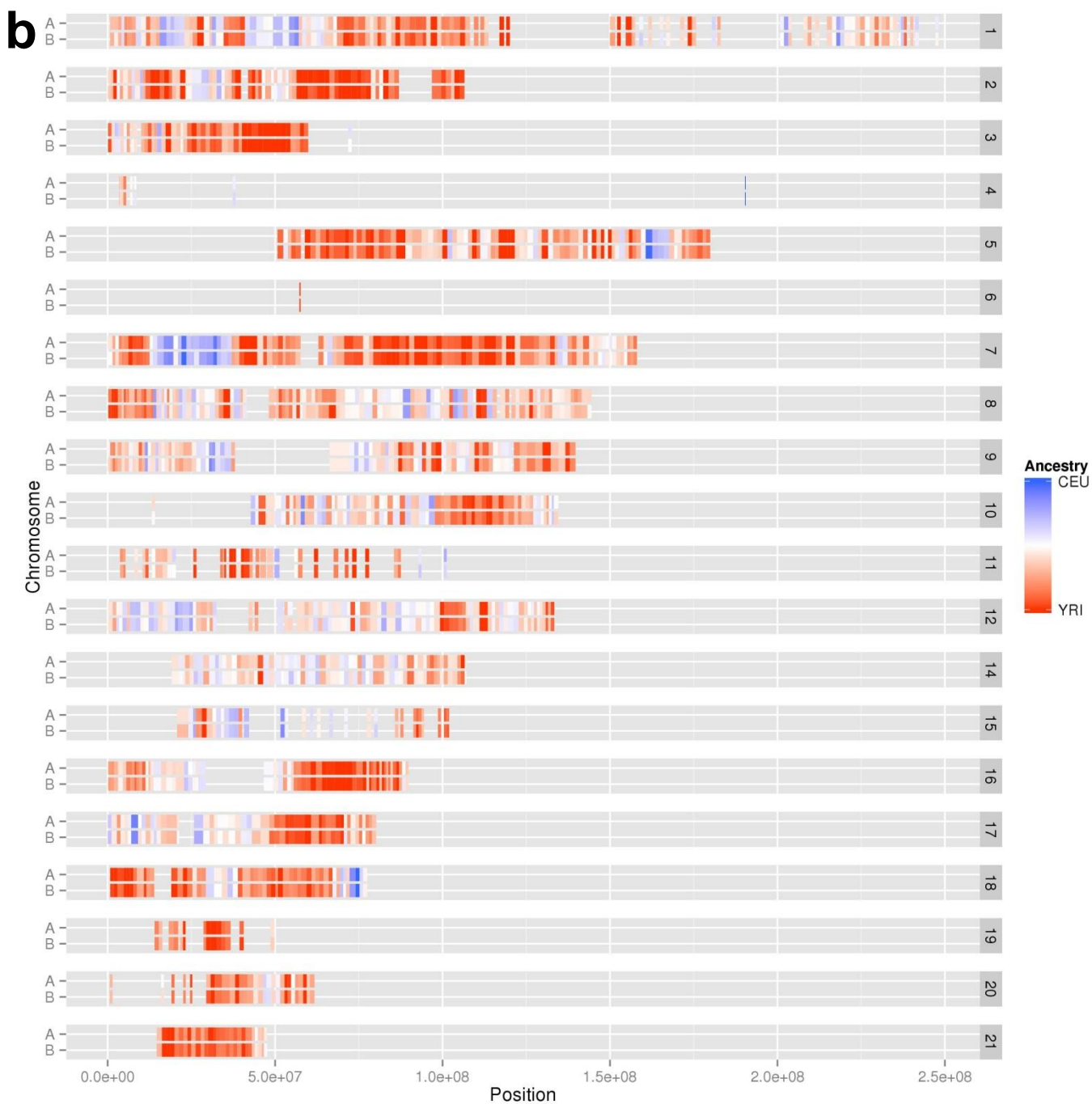


**Figure S18 | Population-based haplotype analysis.**

**a.** Histogram of windowed scores based upon phased sites' population allele frequencies in CEU vs YRI individuals (from the 1000 Genomes Project). Red and green lines indicate density from a two-component mixture model fit. **b.** Randomization test. Histogram of windowed scores, identical to **a**, except that the phase is randomized between each successive pair of 1000 Genomes variants. **c.** Counts of novel variants (non-1000 Genomes Project) for windows ranked as in **a**. (windows with more CEU-like alleles to the left, more YRI-like alleles to the right). More highly YRI-like haplotype blocks on average contain more novel variants. **d.** Randomization test. Counts of novel variants in each window, identical to **c**, except that the phase is randomized as in **b**.

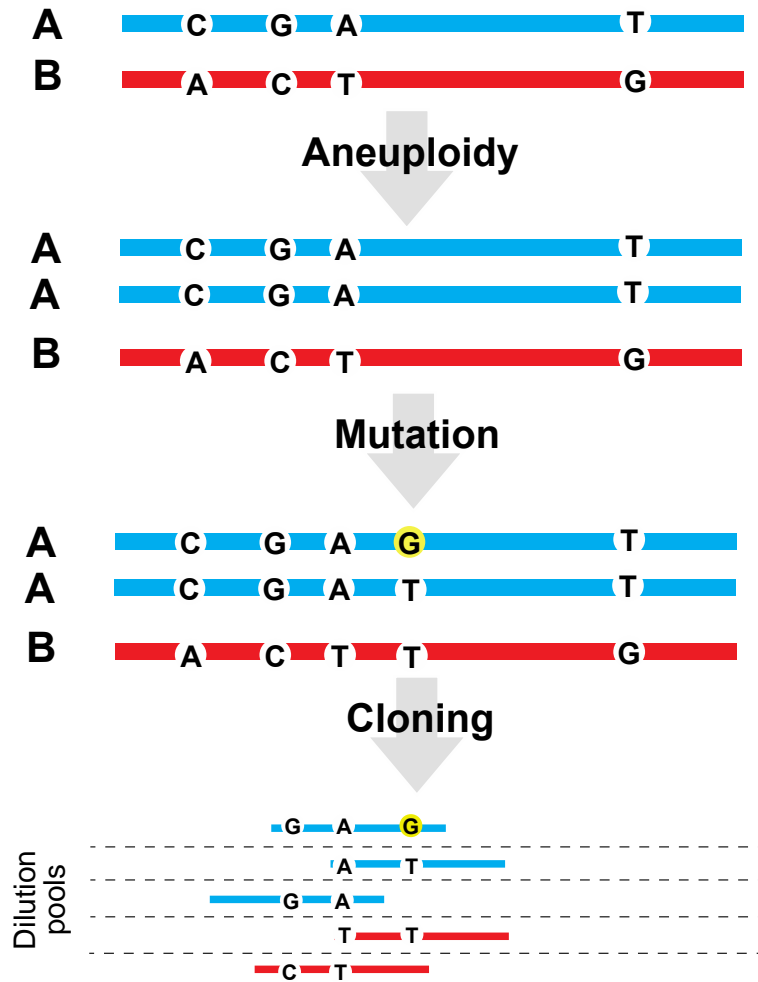
**a**





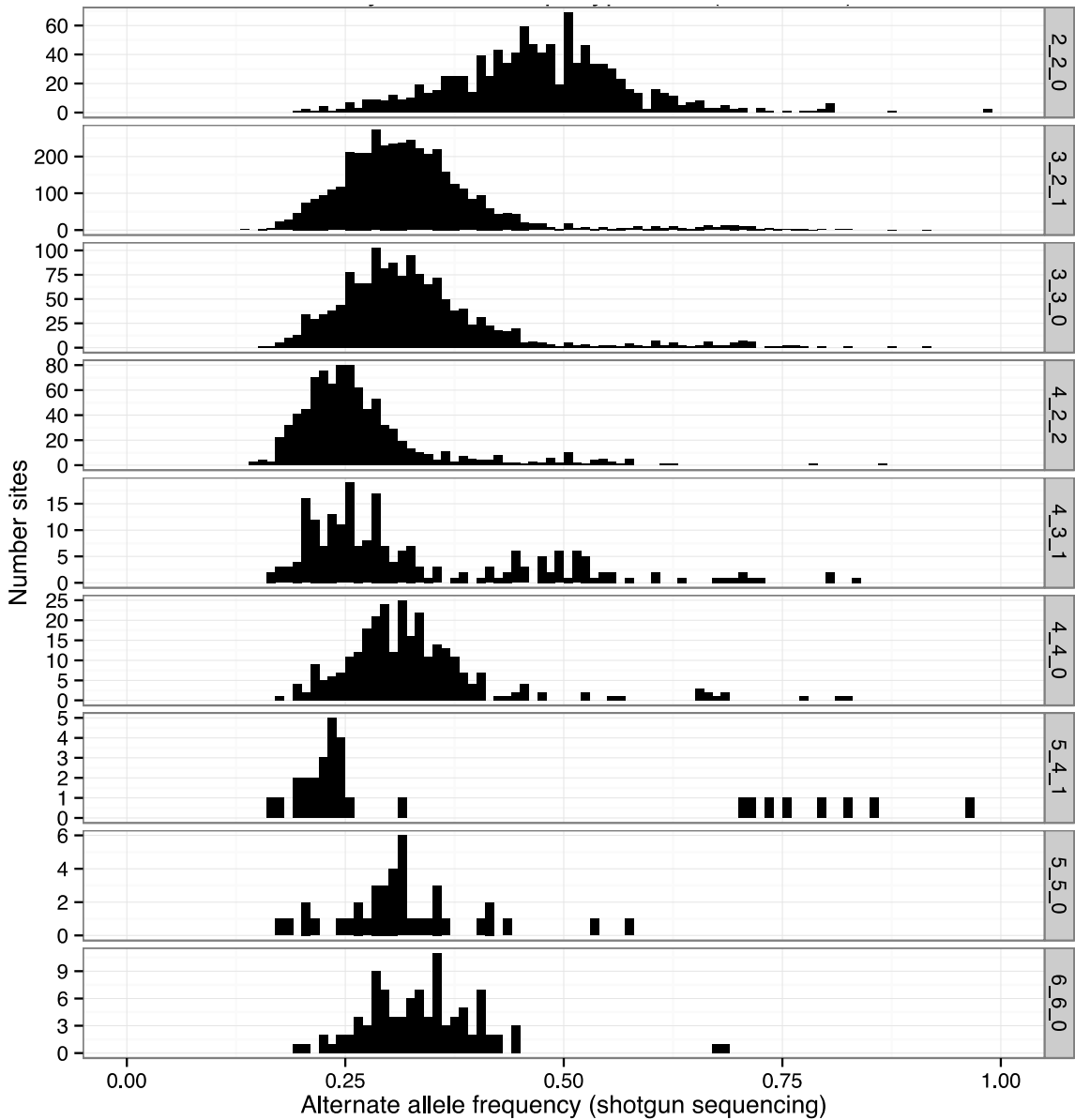
**Figure S19 | Haplotype-based local inference of genetic ancestry.**

- a.** Predicted genetic ancestry is shown for haplotype windows, using scores of allele frequency to CEU or YRI populations and colored by the ancestry similarity (Blue = CEU, Red = YRI). Windows in LOH regions, in haplotype scaffolds with insufficient numbers of phased variants (fewer than 1,000 variants 1000 Genomes Project variants), are not shown. Regions of balanced copy number shown by black boxes were excluded because haplotype imbalance could not be used to create long scaffolds.
- b.** Randomization test. Windows are painted as in **a**, except that the phase is randomized between each successive pair of 1000 Genomes variants.



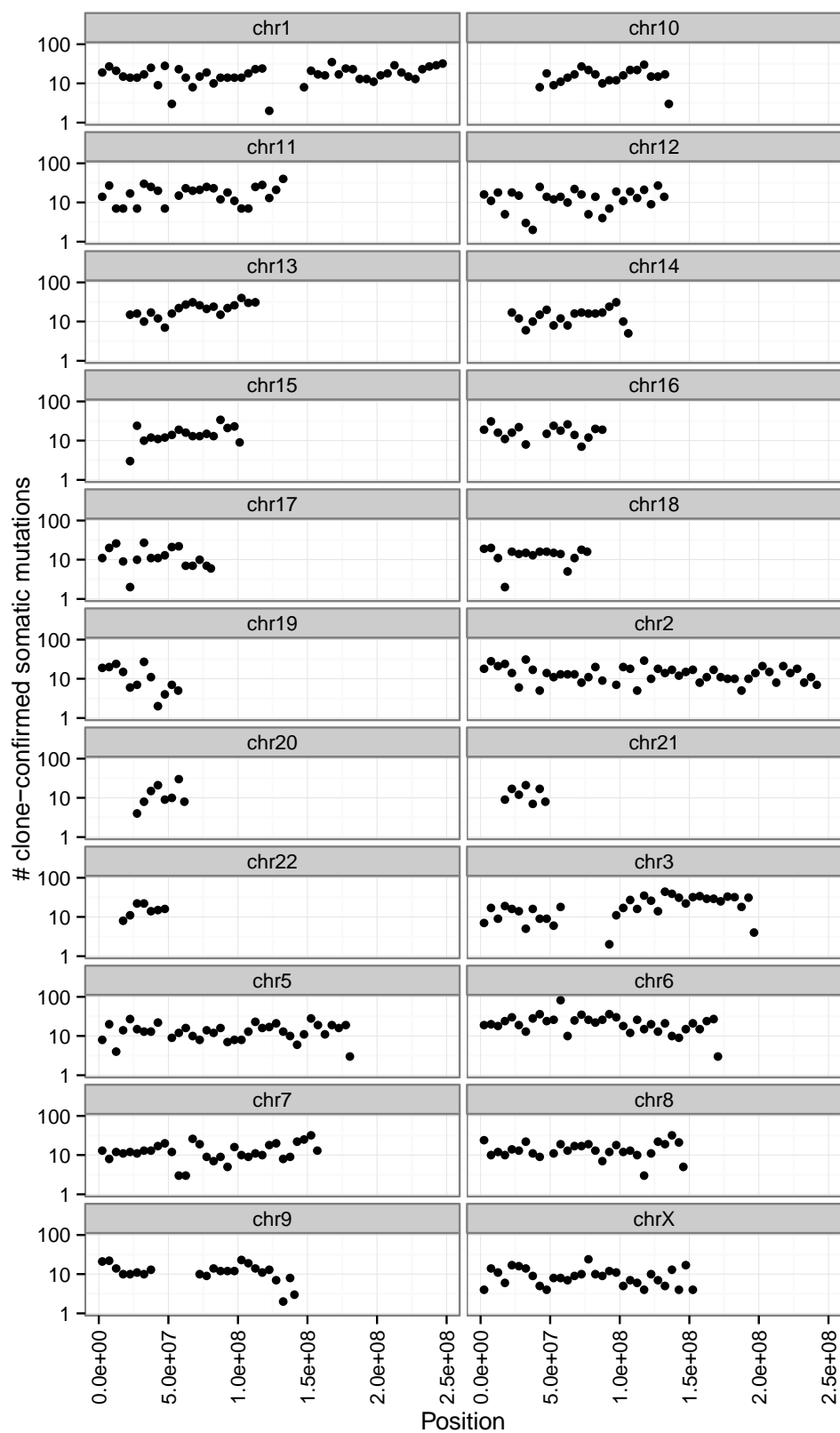
**Figure S20 | Post-aneuploidy mutation analysis.**

Schematic of validation process for somatic, post-aneuploidy mutations by large insert clone pool sequencing. Mutations arising after duplication of a germline haplotype (blue) are confirmed by the presence of both the mutant allele (yellow, “G”) as well as the reference allele (T) in separate clones derived from the duplicated haplotype.



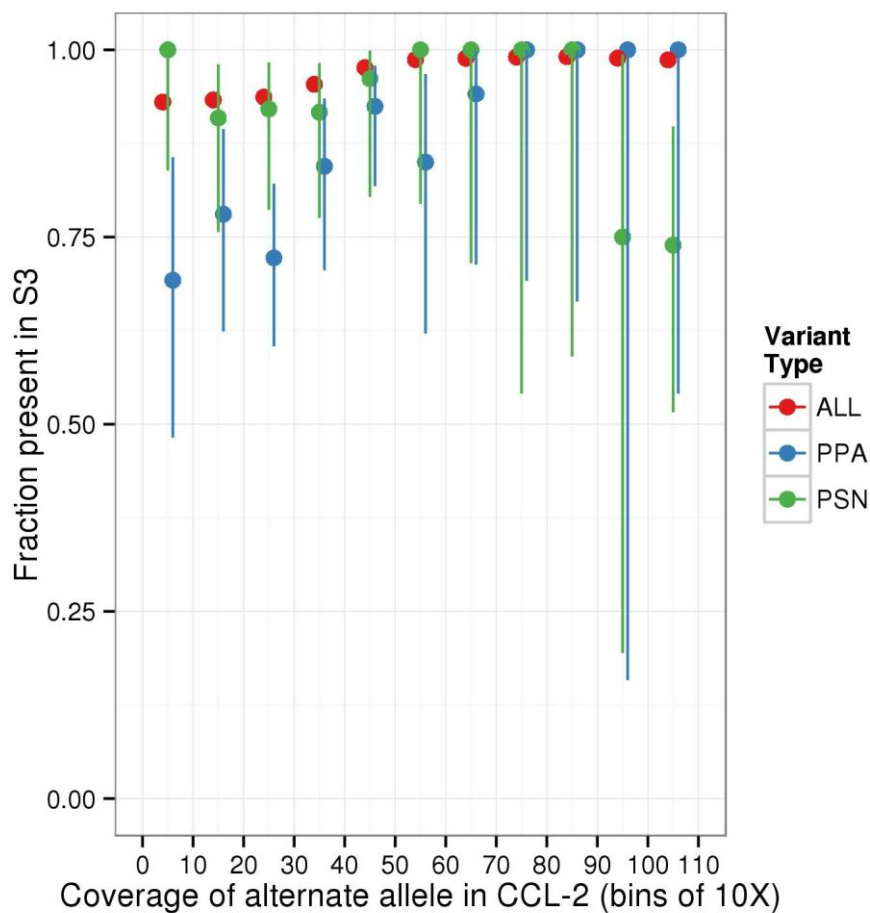
**Figure S21 | Somatic mutation allele frequencies.**

Histograms of allele frequency within shotgun data of clone-validated somatic mutations, split by haplotype-resolved copy number (HRCN) state. Regions with HRCN of 5:3:2 and 6:4:2 were omitted because there were few sites (each  $\leq 10$ ).



**Figure S22 | Somatic mutation counts.**

Count of somatic mutations per 5 Mbp window along each chromosome.



**Figure S23 | Private alleles shared between HeLa CCL-2 and S3.**

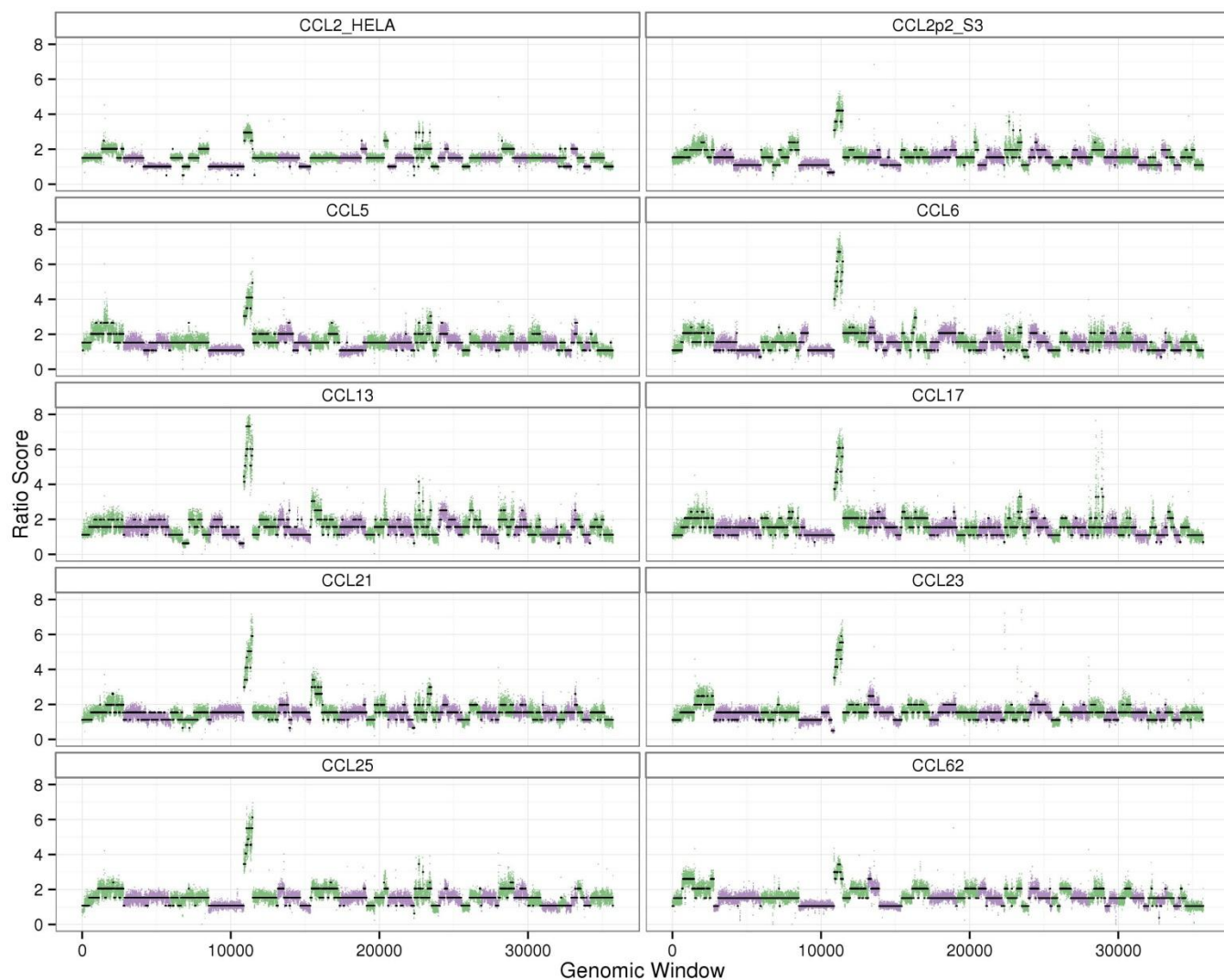
The fraction of private SNVs (not found in the 1000 Genomes Project) from HeLa CCL-2 that are also observed in S3 is shown, binned by the number of reads supporting the alternate allele in CCL-2. The fraction of shared alleles is shown for different categories of sites: all private sites in CCL-2 (Red, “ALL”), private protein-altering variants in CCL-2 (Blue, “PPA”) and private coding synonymous variants in CCL-2 (Green, “PSN”). Variant alleles supported by >100 reads in CCL-2 were grouped into the “100+” bin.





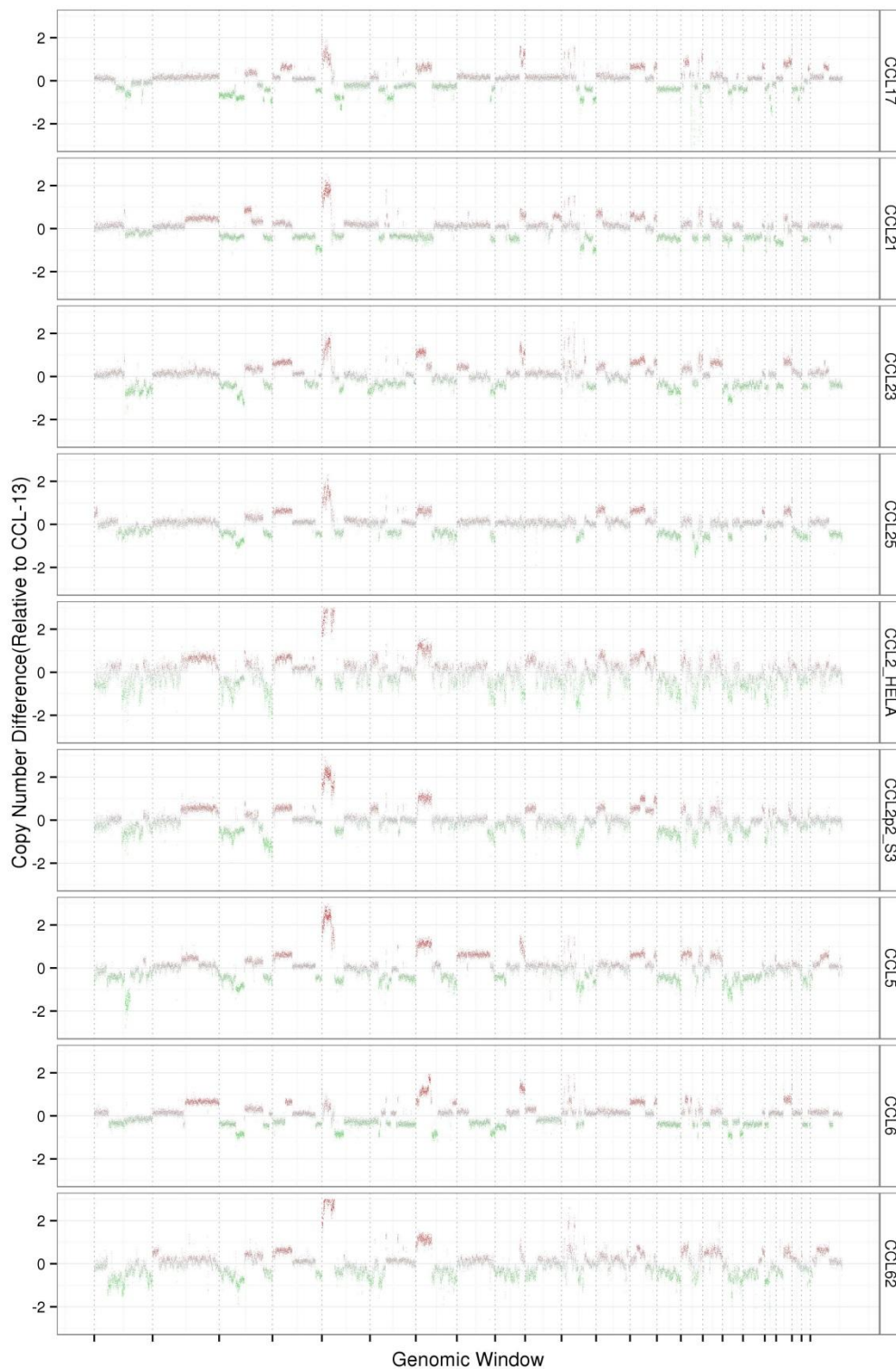
**Figure S24 | HeLa S3 high resolution copy number calls.**

Copy number ratios versus control genomes are plotted within high-resolution SUNK windows (green dots, each window size ~1.5 kb), with predicted copy number state overlaid (black dots).



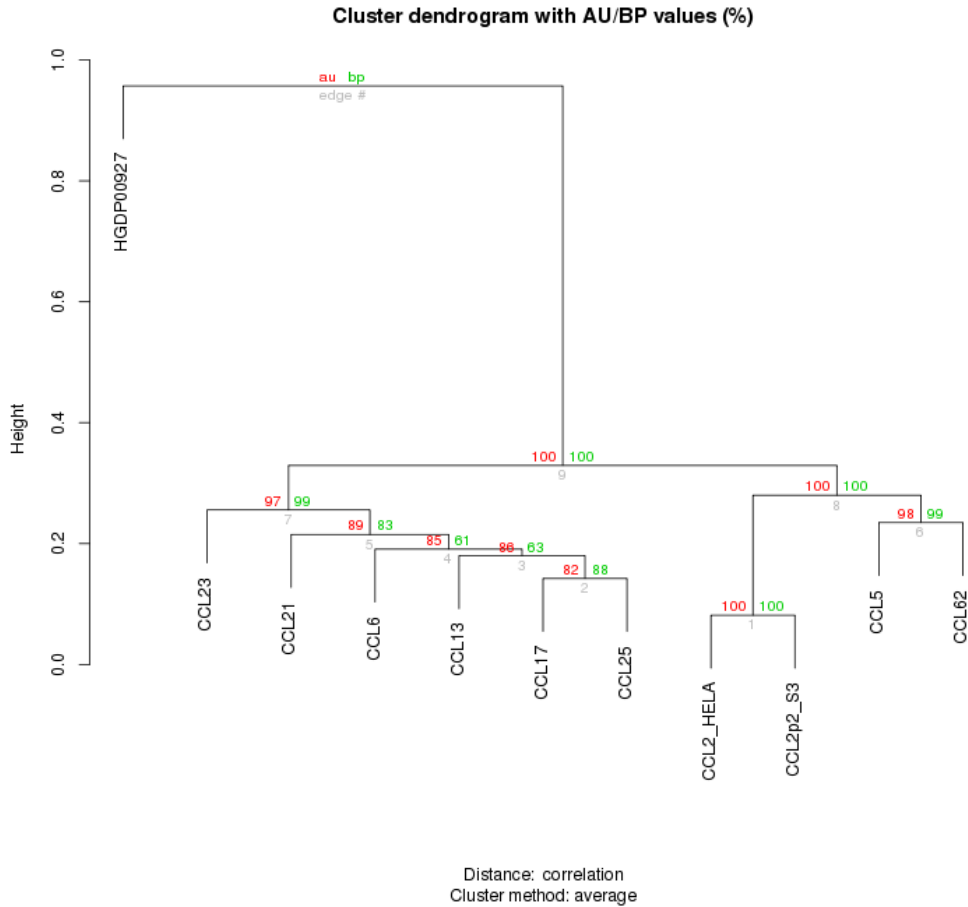
**Figure S25 | Copy number profiles for 10 HeLa strains.**

Copy number ratios versus control genomes are plotted within high-resolution SUNK windows for HeLa CCL-2, HeLa S3, and eight additional HeLa strains (green and purple dots, alternating by chromosome, window contains 500 unique 30mers), with predicted copy number state overlaid (black dots).



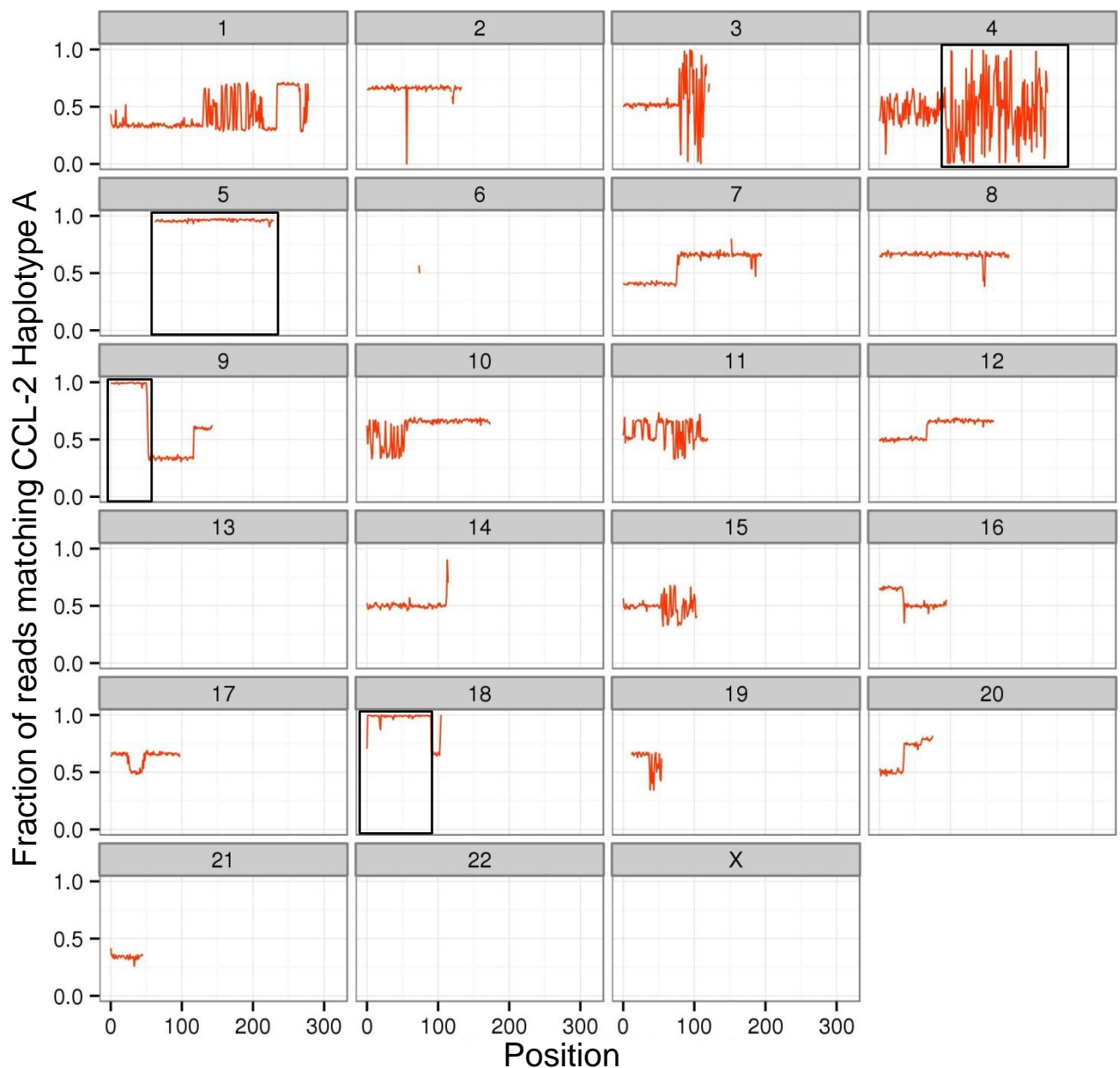
**Figure S26 | Comparison of read depth profiles in HeLa strains.**

Copy number differences across low-resolution SUNK windows relative to HeLa CCL-13 were plotted for HeLa CCL-2, S3, and 7 additional strains. Note: Increased values indicate increased copy number in CCL-13 compared to alternate strain.



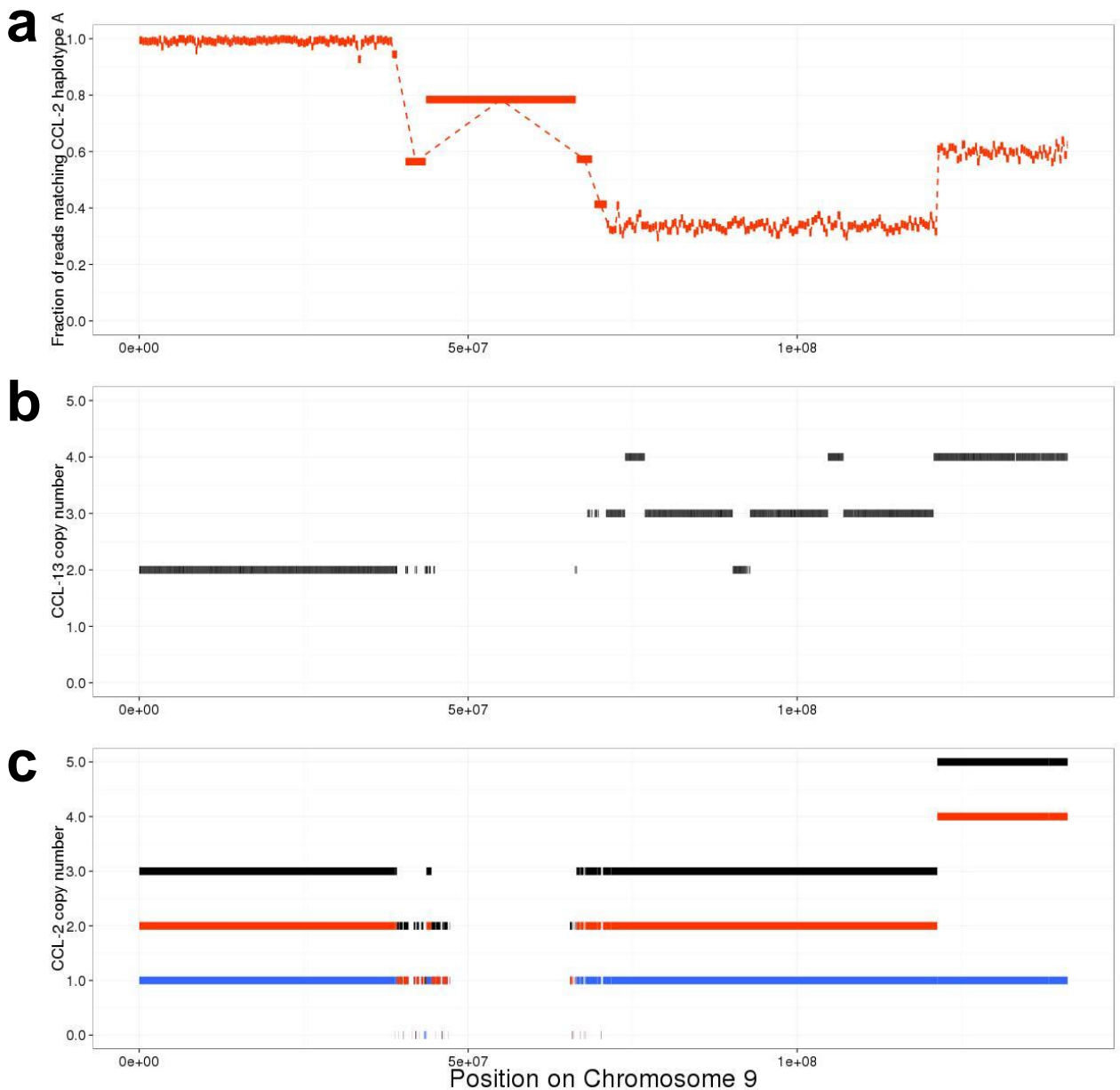
**Figure S27 | Clustergram of 10 HeLa strains based on copy number profile similarity.**

Copy number scores were averaged within large windows (~1 Mbp) for 10 HeLa strains as well as an outgroup control genome (HGDP00927). Scores were clustered in (R package 'pvclust') with 1000 bootstrap iterations. “au” values correspond to “Approximately Unbiased” scoring that is computed by multiscale bootstrap resampling while the “bp” value corresponds to “Bootstrap Probability”, or standard bootstrap scoring. Due to batch differences in library preparation, comparison with HeLa CCL-2, HeLa S3 and the HGDP outgroup is much less reliable. It is important to note that this dendrogram is not necessarily the actual phylogeny and simply represents the similarity between marker chromosome / copy number subsets for the individual strains.



**Figure S28 | Regions of LOH in HeLa CCL-13 by comparison to CCL-2 haplotypes.**

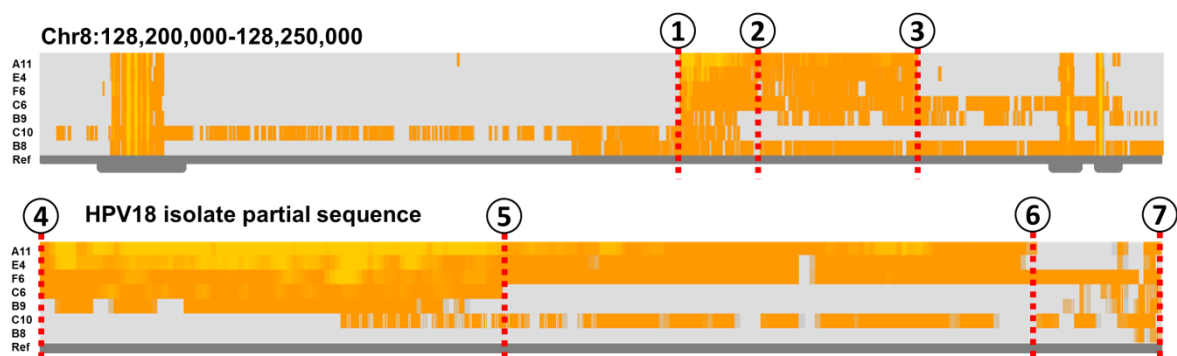
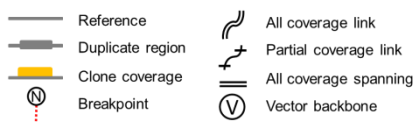
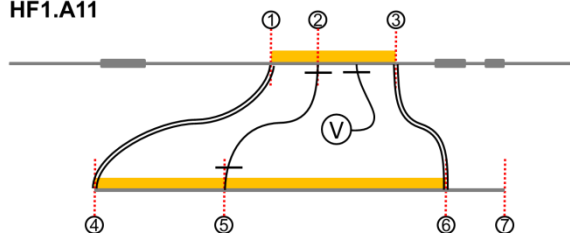
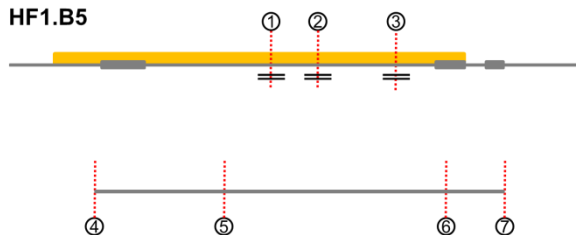
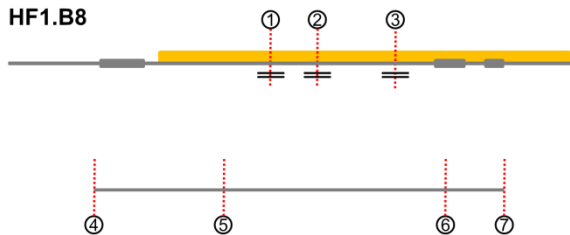
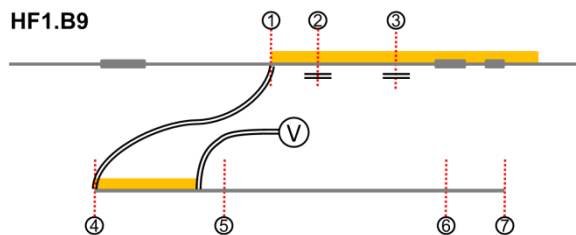
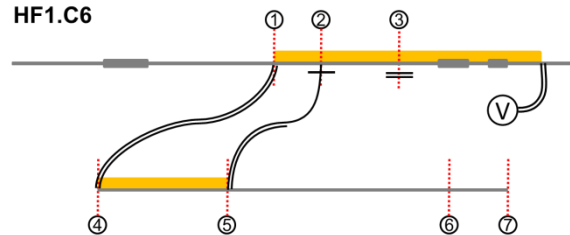
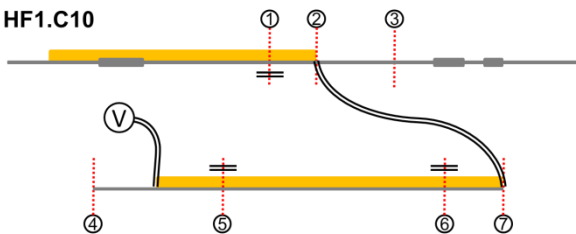
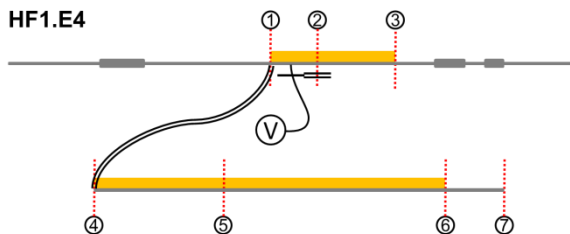
Shown in windows (mean ~800 kbp) across each chromosome are the fraction of reads matching the allele phased to haplotype A in HeLa CCL-2. LOH in CCL-13 (but not CCL-2) manifests as long stretches where shotgun reads from CCL-13 (mean depth 4.0X) exclusively match CCL-2 haplotype A (y value = 1) or haplotype B (y value=0). A total of NNN Mbp of LOH regions were detected in CCL-13 (highlighted by shaded bars). Regions lacking haplotype scaffolds in CCL-2 (e.g., in LOH or in regions of balanced copy number in CCL-2) were omitted. Black boxes indicate predicted regions of LOH.



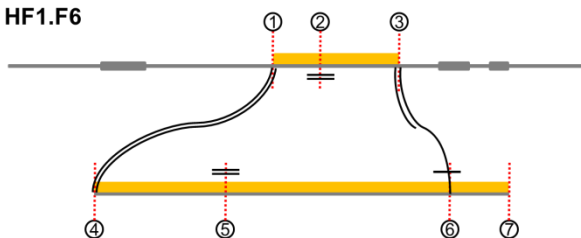
**Figure S29 | Copy-number loss and LOH on chromosome 9 in HeLa CCL-13.**

**a.** LOH on chromosome 9 in HeLa-CCL13, as detected by a shift towards CCL-2 haplotype A alleles was accompanied by reduction of copy number to 2 in CCL-13 in the affected region shown in **b**. relative to copy in HeLa CCL-2 shown in **c**. (Black = total copy number, Red = haplotype A copy number, Blue = haplotype B copy number).

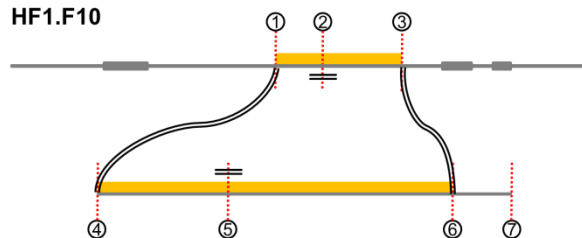


**a****b****Legend****HF1.A11****HF1.B5****HF1.B8****HF1.B9****HF1.C6****HF1.C10****HF1.E4**

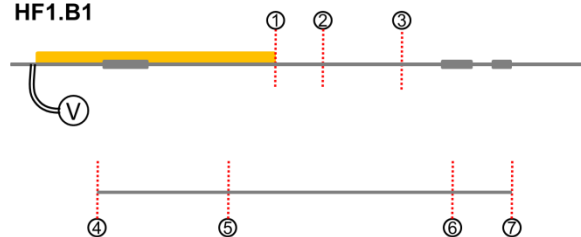
HF1.F6



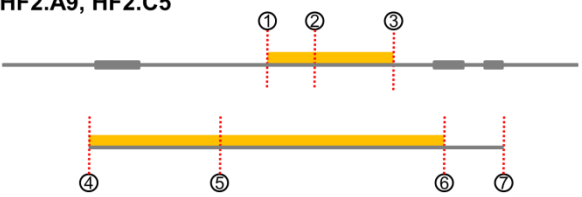
HF1.F10



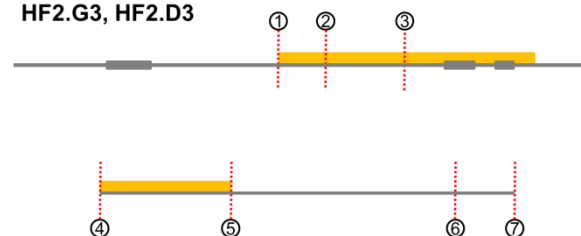
HF1.B1



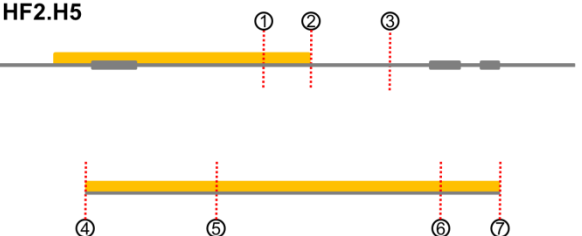
HF2.H10, HF2.B1, HF2.G9, HF2.H11, HF2.A5, HF2.A9, HF2.C5



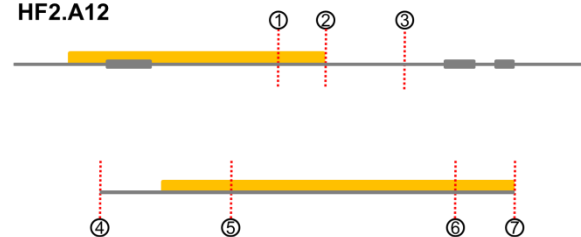
HF2.G3, HF2.D3



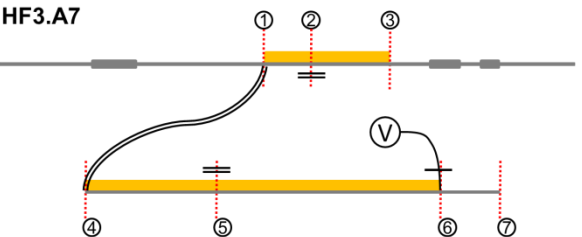
HF2.H5



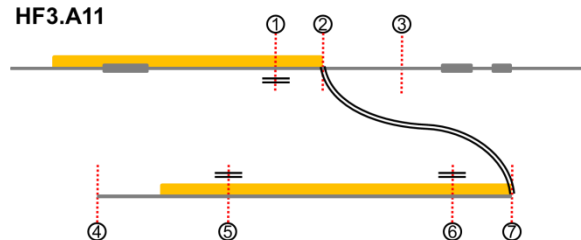
HF2.A12



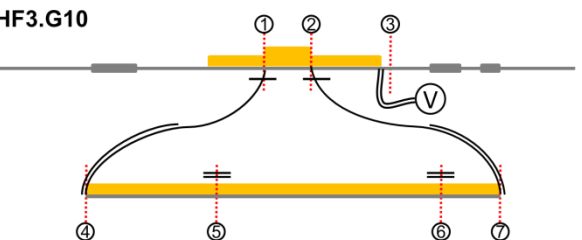
HF3.A7



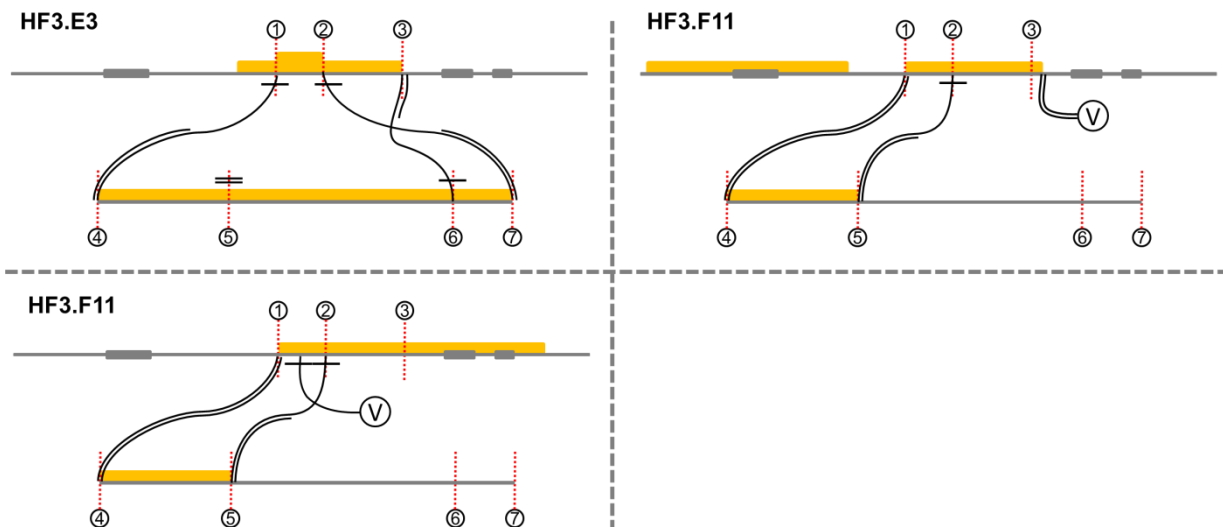
HF3.A11



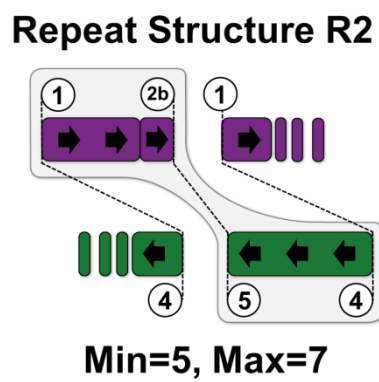
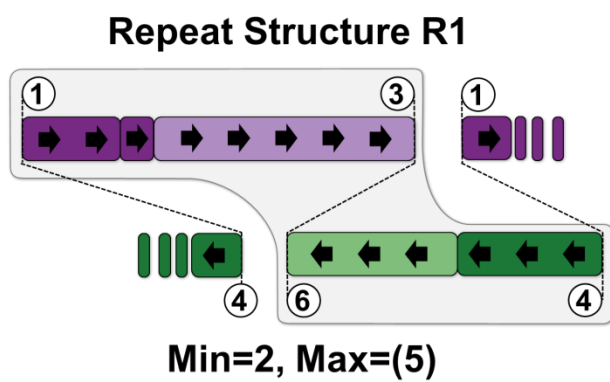
HF3.G10





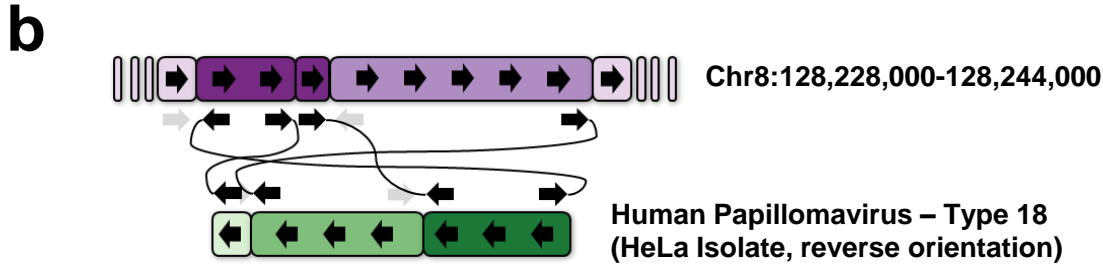
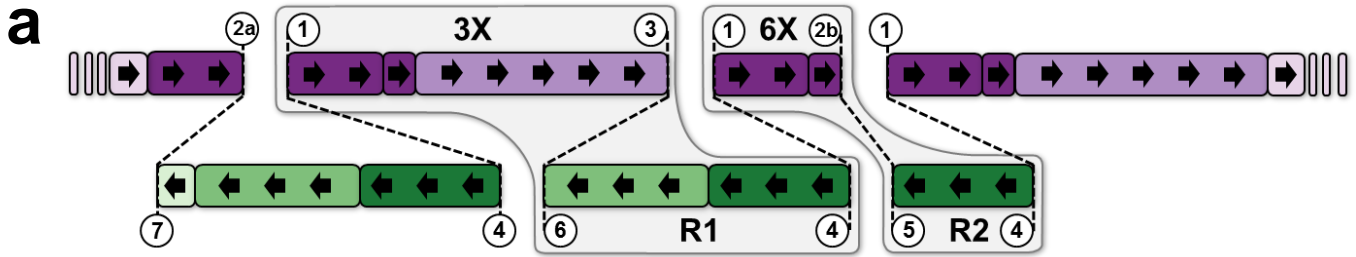


**C**



### **Figure S30 | Structure of the HPV-18 integration locus.**

**a.** Heat maps showing coverage from fosmid clone pools across the insertion site flanking region (upper) or the HPV-18 partial sequence (lower). Circled numbers correspond to breakpoints in which discordantly mapping read pairs were found that link the HPV-18 and chromosome 8 references. **b.** Diagrams of individual clones' coverage across the integration site and HPV-18. Yellow bars indicate coverage in the region, double black lines parallel to the reference indicate all read pairs are concordant across the breakpoint; single black links with a single line parallel to the reference indicate some read pairs are concordant across the breakpoint and others correspond to the link; double black links indicate that all read pairs support the link; links to a circled “V” indicate that read pairs are present at that location that link to the backbone vector sequence and therefore mark the end of a clone's coverage. **c.** Proposed structure of repeat units based on fosmid coverage profiles. Repeat R1 has a minimum of 2 copies, as breakpoint 2b-5 is never observed in fosmids containing breakpoint 3-6. The observed coverage profile over repeat 1 indicates a maximum of ~5 copies. Repeat R2 has a minimum of 5 tandem repeats, as the HPV region from breakpoint 5 to breakpoint 6 is never observed in clones entering the repeat from the centromeric end of the region, and a maximum of 7 due to never observing fosmids solely containing the chromosome 8 region from breakpoint 1 to 2b and HPV region from breakpoint 4 to breakpoint 5.



**c**

**Breakpoint 2a-7 Assembled**

GTT ATT ACA CAG CTA TCA GAG CAA GAG GGA GGT TAG TAA AAG CTG GTG GAC CTT AAA GTT TCT CTA CTT TTG CAA  
GTG TAA AAA CTG GGG TAA AGA TAG AGT TTT GTT TTT CCT CGG TTT TGT ATG CAC TTT GTG CAA GGC CTT GTA GGG  
CCA TTT GCA GTT CAA TAG CTT TAT GTG CTT TAC TTT TTG AAA TGT TAT AGG CTG GCA CCA CCT GGT GGT TTA ATG  
TCT GTA TGC CAT GTT CCC TTG CTG CAA AGA ATA TTG CAT TTT CCC AAC GTA TTA GTT GCC AAT ACT GTA TTT GGC

**Breakpoint 1-4 Assembled**

GAA ACC TTA GGA ATA TCC TGC TTA TTG CCA CCA CCT GCA GGA ACC CTA AAA TAT GGA TTA CCA ACA GTT AAT AAC  
CAG ACA AAA ACT TTA ATA ATA TTT GTC AAA TGC CAA ATC GGA GTC CAA AGC CAT TGT CCA TTT TAA GAA AAT CAT  
CTG ACT TAA CAT CAC TAC TGC TTT TCA AGA GAG CAT CAT GCC CAT TTC ACA GAA GAG AAA ATT TGG CCT CAT ACT  
CCT CAG TCT CCA TGT TTT AGC TTA GAT ATT GTT TCC TCC AGG CGG CCC TCC TTG ACC TTC CAA TTC TGG TTA AAT  
TGC TCT TTT TCT GAG TTC TCA TTA CTT TAC TGA TTT TAT ATA TGT GTG TGT GTA TAT ATA TAT ATA CAC ACA CAC  
ACA CAC ACA CAT ATA TAT ACA CAC ACA TAT ATA TAT ATG TTG TGC CTA GCA AGT GTA TGA CAC AAA ATA CCC ATA  
AAT TGA ATG AAT GAA TGA ATT AAT AAA GAA ATG AAT AAC TTA CCC AAC CTG GTA AGT GGC AGG GCT GGC CAG GTC  
AGT GCA ACT TCA AAG TCG ATG TTG TCA GTG AAT GCT CCA GAT GGA TTG CAG AGA AGA CC

**Breakpoint 3-6 Assembled**

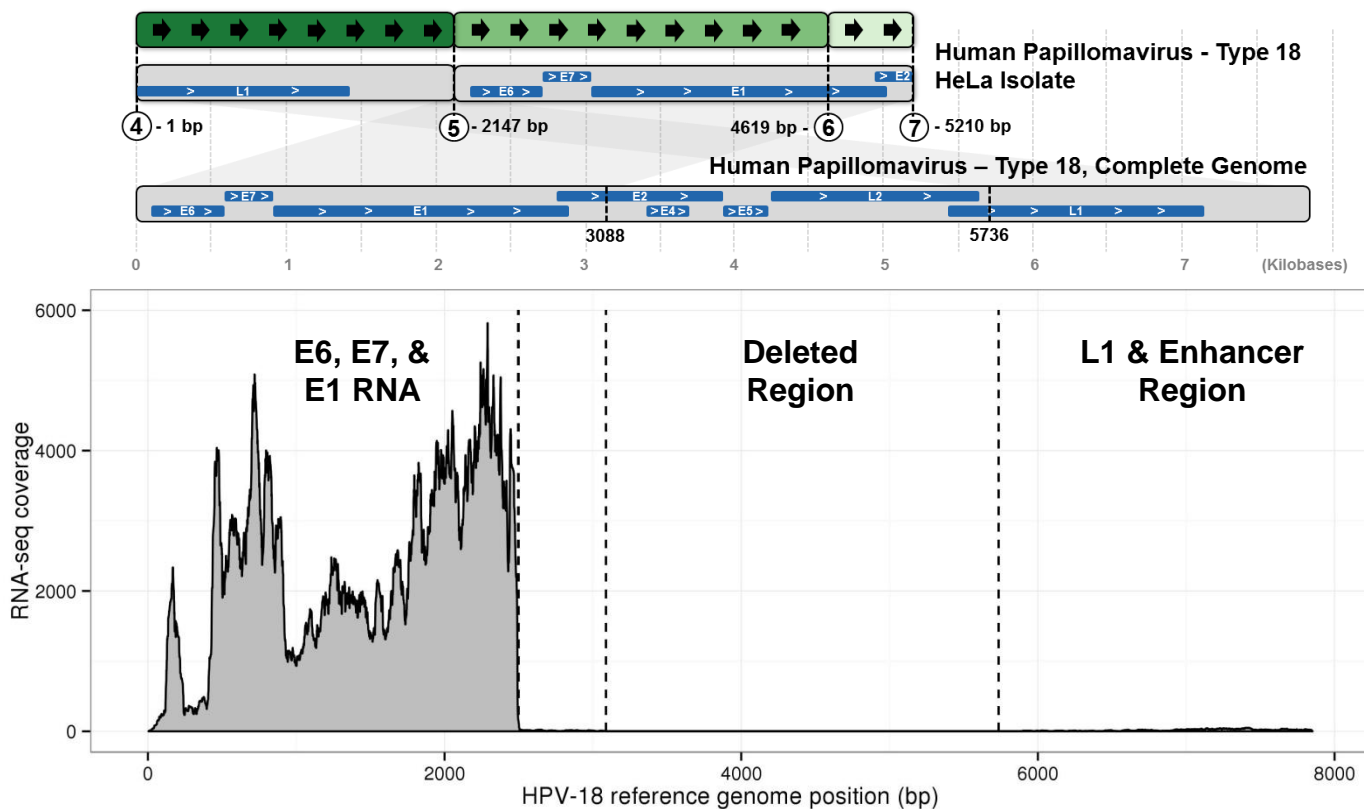
TCT TCT ATG AGC TTC GTC AAG TCA TTT AAG CTT GGT ACC CGT CAG TTT CCT CAT CTG AAA ACT GAG AAA AGT TGT  
TTC AAA TTG TCT AAG TCC ATT CCA GCT TGA TCA TAC TAG CAT CTT ATG TGC AGC TTC TTA AAG TCC AGC TCA CAC  
CTC TGT CAA CTC CCT GTA TAA TAT GAC TTC CAA AAA AAC ACC TGT GGT TTG GTT ATA CAT ATA TAT GGA CAT ATA  
TAT GTT ATA ACA TGG CCA CCT TAG TAT CTG TTA ACG GTT CCA ACC AAA AAT GAC TAG TGG AAT TCA CAA ATG ATA  
TTA CTG CTC CTT GTA TAA AGT GTA TAA AAC TCA TTC CAA AAT ATG ATT TTC CTG TAT TTG CTG GTC CAC AAA ATA  
CTA AAC

**Breakpoint 2b-5 Assembled**

AGG AAC AAA GGA ATC GAG GGA GGA AGG GAA GAA AAA ATG AGA AAA ACC ATA AGG CCA GGC GCG GTA GCT CAC GCC  
AGT AAT CCT AAT ACT TTG GGA AGC TGA GGC GGG TGG GCG GAC CAC GAA GTC AGG AGT TCG AGA CCA GCC TGA CCA  
ATA TGG CAA AAC CCC ATC TCT ACT AAA AAT CCC AAA AAA AAA AAA AAA AAA AAG TTA GCC GGG TAT GGT  
GGC ACG TGC CTG TAA TCC CAG CTT CTG GGG AGG CTG AAG CAG GAG AAT TGC TTG AAA CCG GGA GGT GGA GGT TGC  
AGT GAG CCG GGA TCA CAC CAC TGC ACT CCA GCC TGG GTG ACA GAG TGA GAC TCC CTC TCA AAA AAA AAA AAA AAA  
AAA AAA AAA AAA AAA AGA AAA AGG AAA AAG AAA AAA AAG CAA CCA TGA GAC GAG CAA GAA GCT AAG TTT ACT ACA  
ATT GTT AAA AGT ATT AAT GAA AAG TAT AGT ATG TGC TGC CCA ACC TAT TTC G

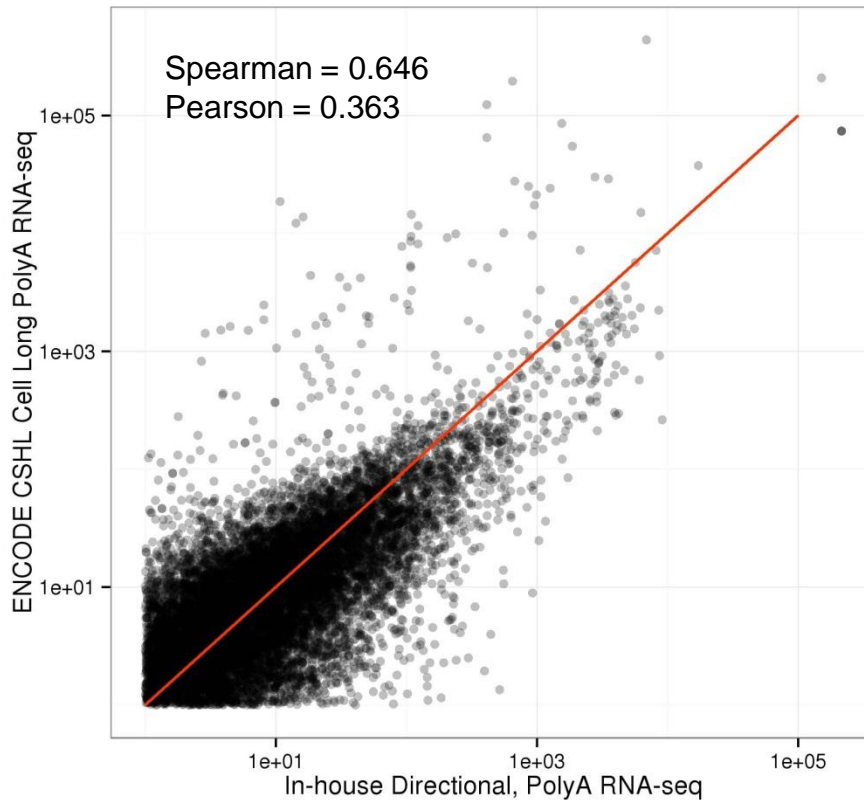
**Figure S31 | Assembly and sequencing of the HPV-18 integration site.**

**a.** Proposed structure of the chromosome 8 locus containing the HPV-18 integration. **b.** Priming sites used to generate amplicons for breakpoint confirmation and assembly. Connecting black arrows indicate successful PCR amplicons and assembled breakpoints, gray arrows indicate additional primer sites that were tested which did not yield products. **c.** Assembled breakpoints performed via shotgun sequencing and assembly of gel-based size selected amplicons. Purple corresponds to human sequence and green to viral sequence, black nucleotides without an underline indicate sequence that share no homology with human or HPV-18 sequence, black nucleotides with double underline indicate sequence micro-homology with both human and HPV-18 sequence, underlined regions in color are the primer sequences used to generate amplicons.



**Figure S32 | HPV-18 RNA-Seq coverage.**

Area chart (bottom panel) represents RNA-Seq level of coverage that reaches nearly 6,000 fold. Above the chart is the diagram of the HPV-18 portion of the integration locus on chromosome 8q24.21 from **Figure 2** for reference.



**Figure S33 | Correlation between RNA-Seq datasets.**

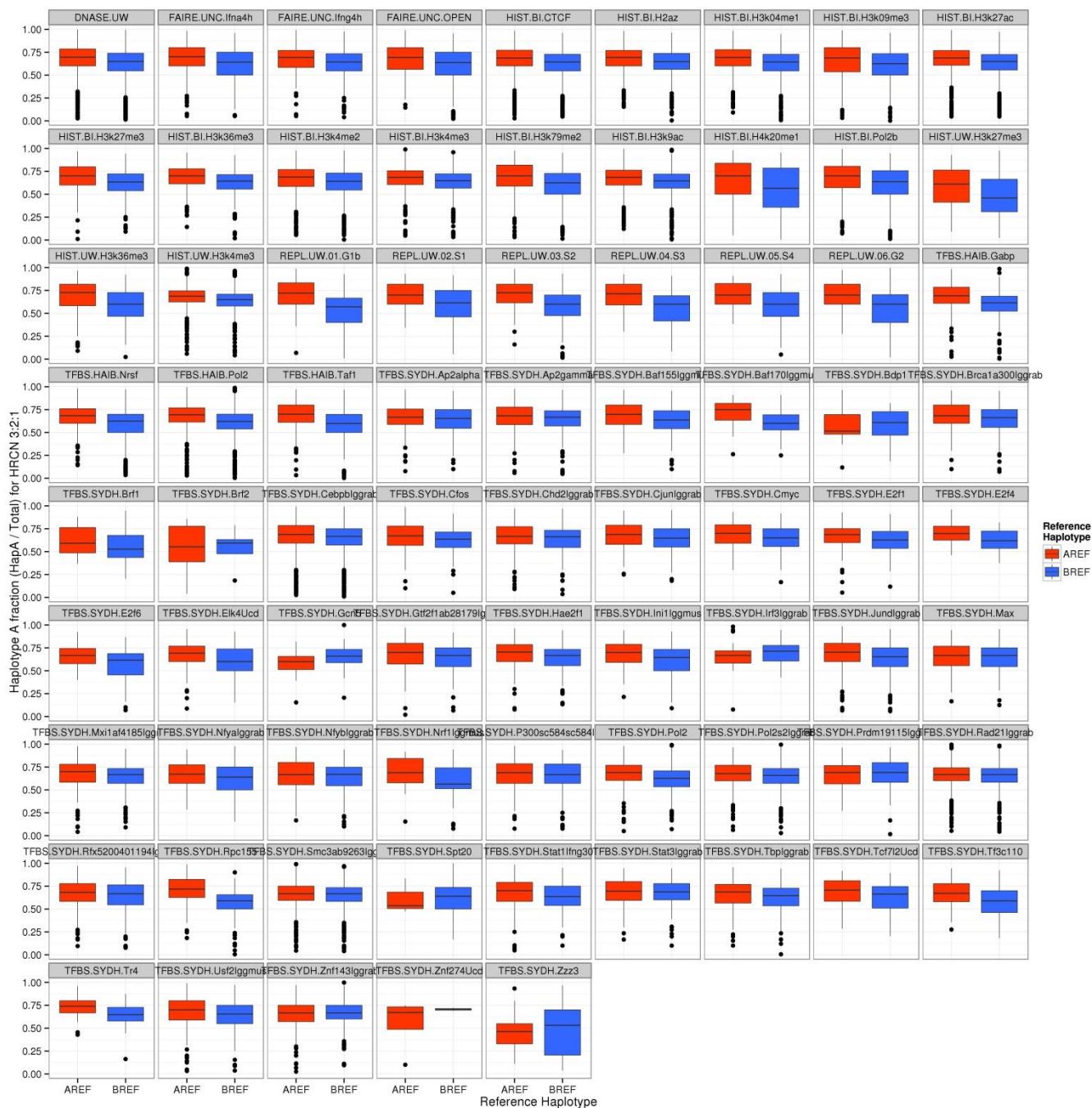
HeLa S3 transcript abundances (reads per kilobase per million reads, RPKM) from ENCODE RNA-Seq (Cold Spring Harbor – Cell long PolyA) were plotted against those our own RNA-Seq data. Each point represents one RefGene-annotated transcript (for transcripts with  $\geq 1$  RPKM). Red line is  $y=x$ .





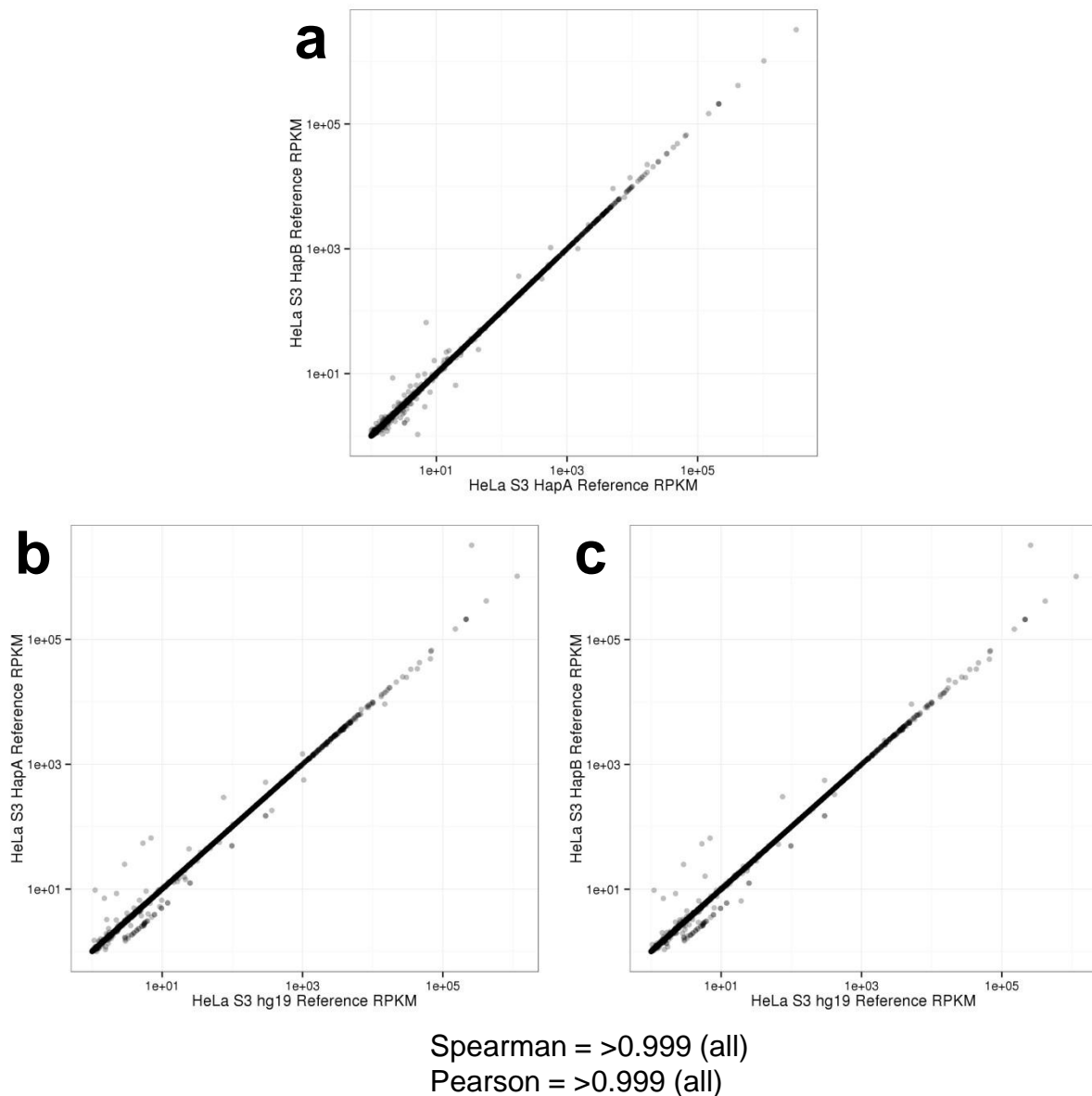






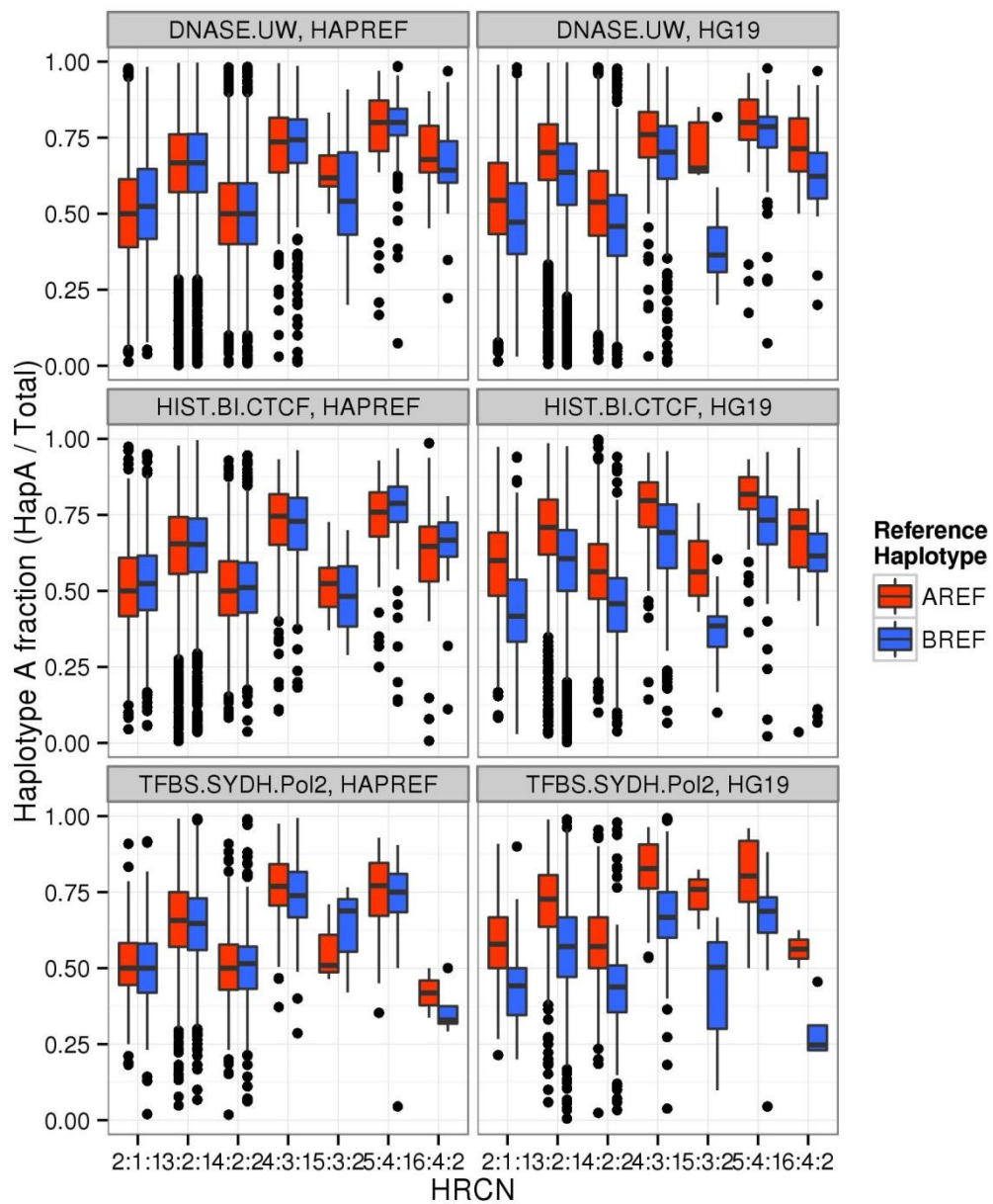
**Figure S36 | Reference bias in ENCODE peaks.**

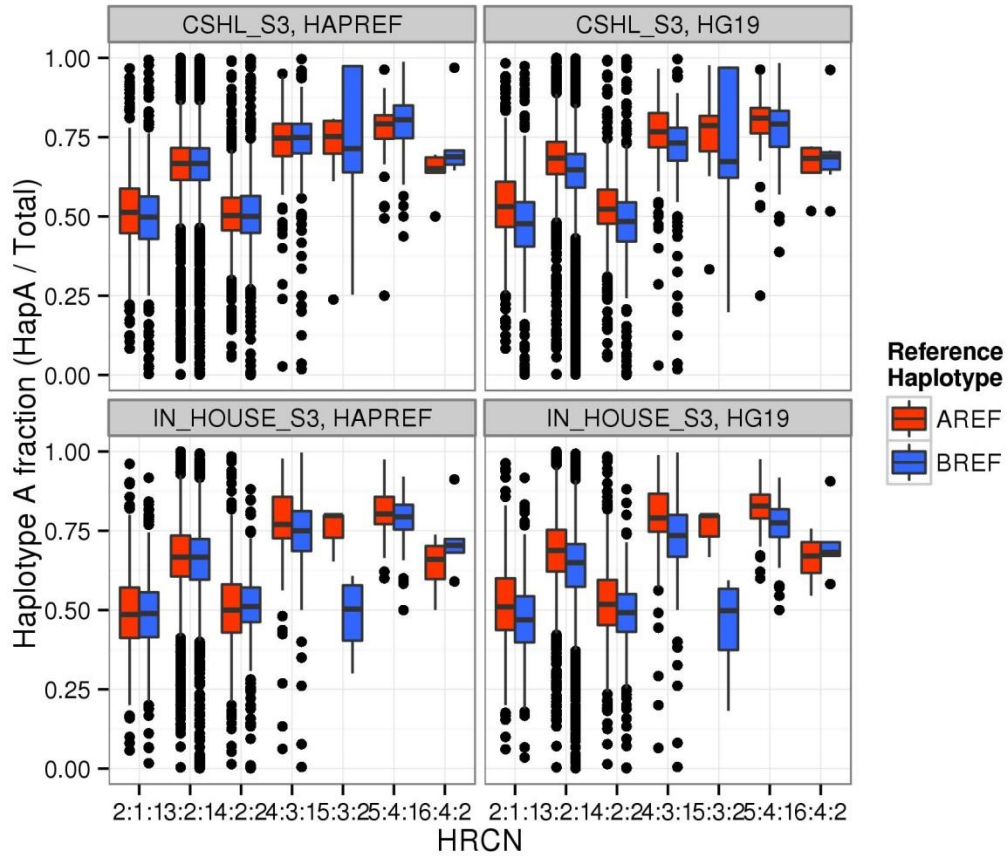
Average degree of reference biases in ENCODE peaks within HRCN 3:2:1 regions are shown as box-and-whisker plots. Red bars represent the haplotype A fractional contribution when the haplotype A allele is the reference base. Blue bars represent haplotype A fractional contribution where the haplotype B allele is the reference base.



**Figure S37 | Minimal impact of reference bias upon transcript quantitation.**

HeLa S3 RNA-Seq reads (this study) were aligned using TopHat (Trapnell *et. al.* (2009)) to the reference genome ("hg19"), as well as to HeLa haplotype-specific reference genomes ("HeLa Haplotype A" and "HeLa Haplotype B"). Transcript abundances were estimated against RefGene annotations using Cufflinks (Roberts *et. al.* (2011)) then compared for all transcripts with an RPKM score  $\geq 1$ . **a.** Comparison between HeLa Haplotype A reference (x-axis) and HeLa Haplotype B reference (y-axis). **b.** Comparison between hg19 and HeLa Haplotype A reference (y-axis). **c.** Comparison between hg19 and HeLa Haplotype B reference (y-axis).

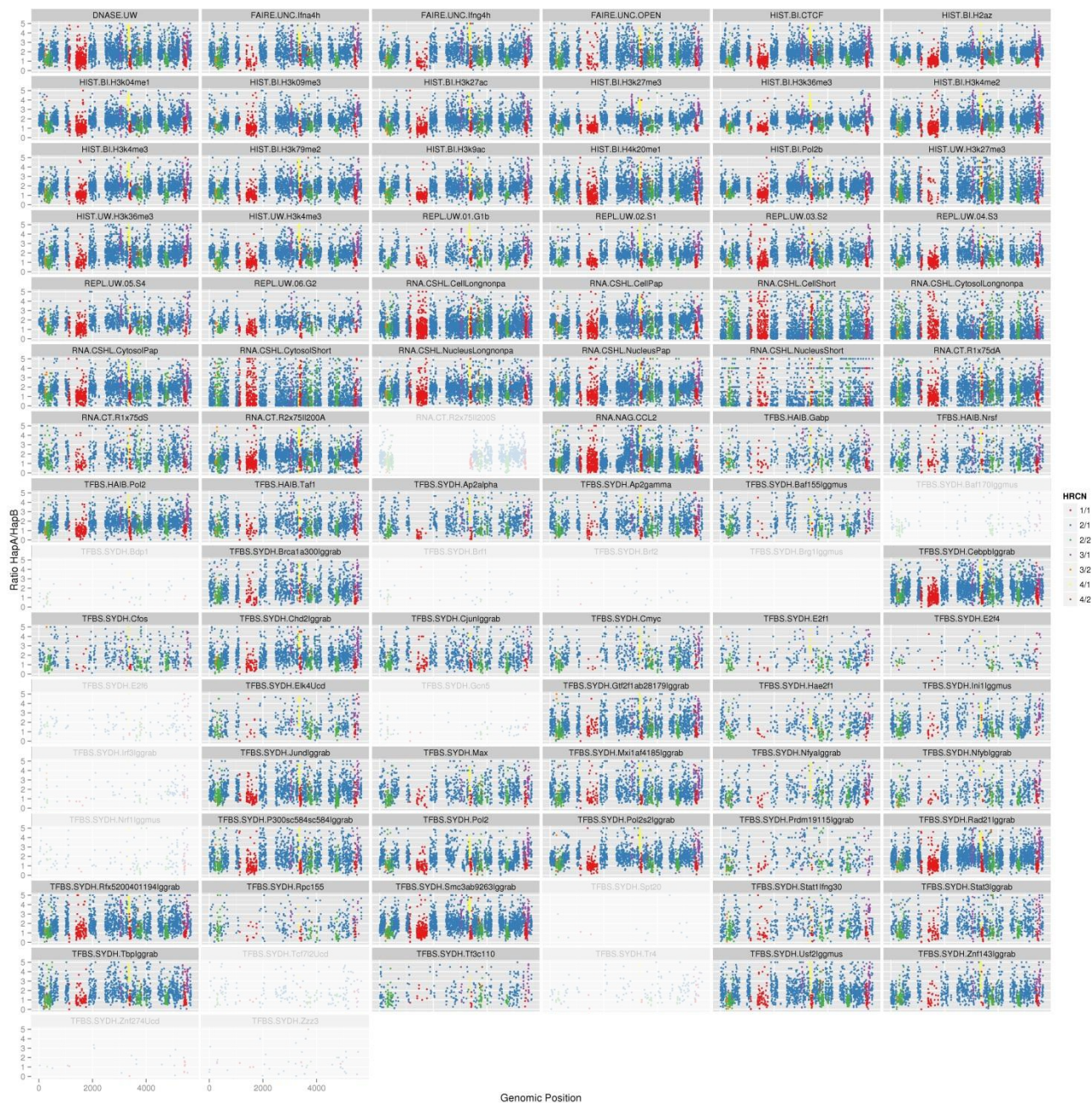




**Figure S38 | Reference bias removal.**

Reference haplotype imbalance for different HRCN classifications for in HeLa when aligning to a HeLa haplotype-resolved reference (left) or hg19 (right). **a.** Reference bias in ChIP-seq peaks. **b.** Reference bias in RNA-Seq. Red bars represent the haplotype A fractional contribution where the haplotype A allele is the reference base. Blue bars represent haplotype A fractional contribution where the haplotype B allele is the reference base. The use of a haplotype-resolved HeLa reference greatly reduced the reference associated bias.

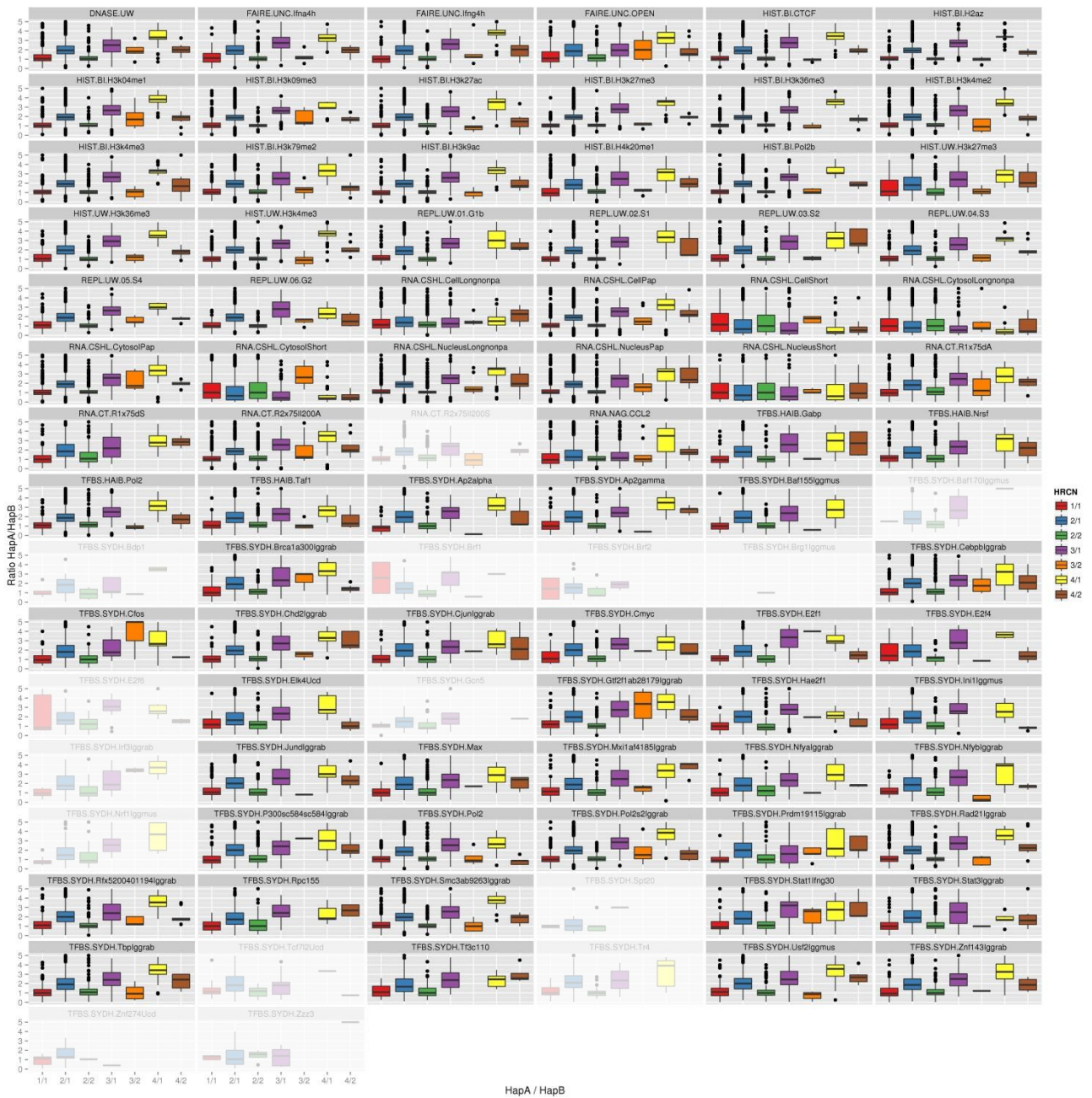




**Figure S39 | Haplotype contributions of phased ENCODE data (windows).**

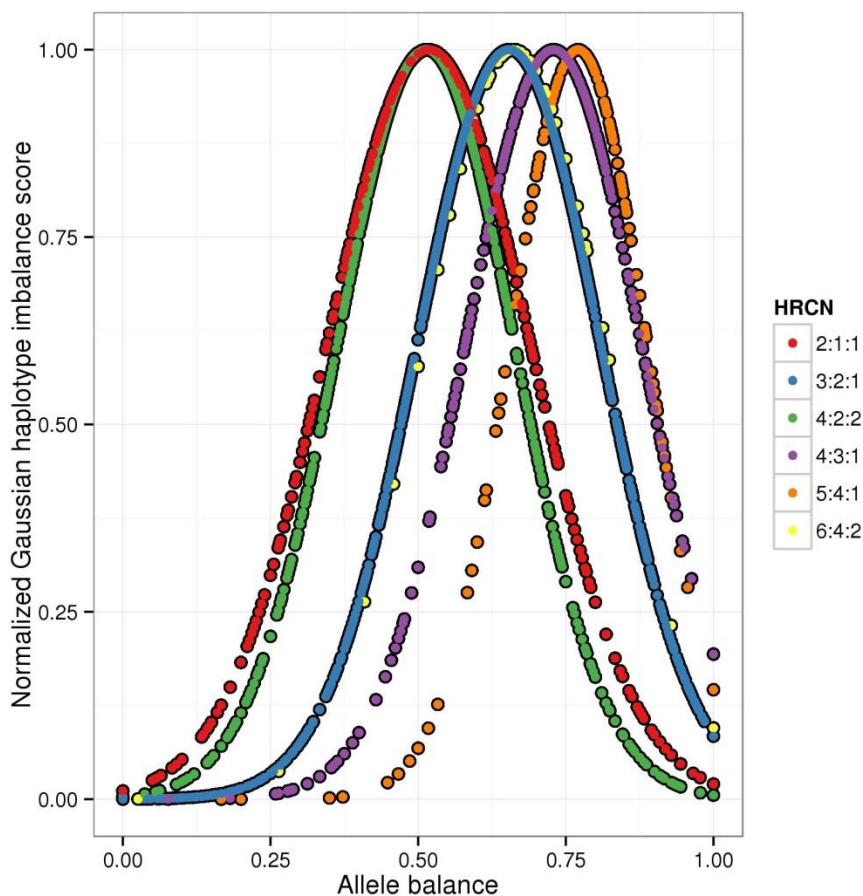
Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows. Each window is color coded by the haplotype A to haplotype B ratio. Dimmed panels indicate data sets with very insufficient numbers of peaks for windowed analysis.





**Figure S40 | Haplotype contributions of phased ENCODE data (box plots).**

Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows shown as box-and-whisker plots. Shaded out panels indicate data sets with very low peak counts and thus can not be reliably analyzed.

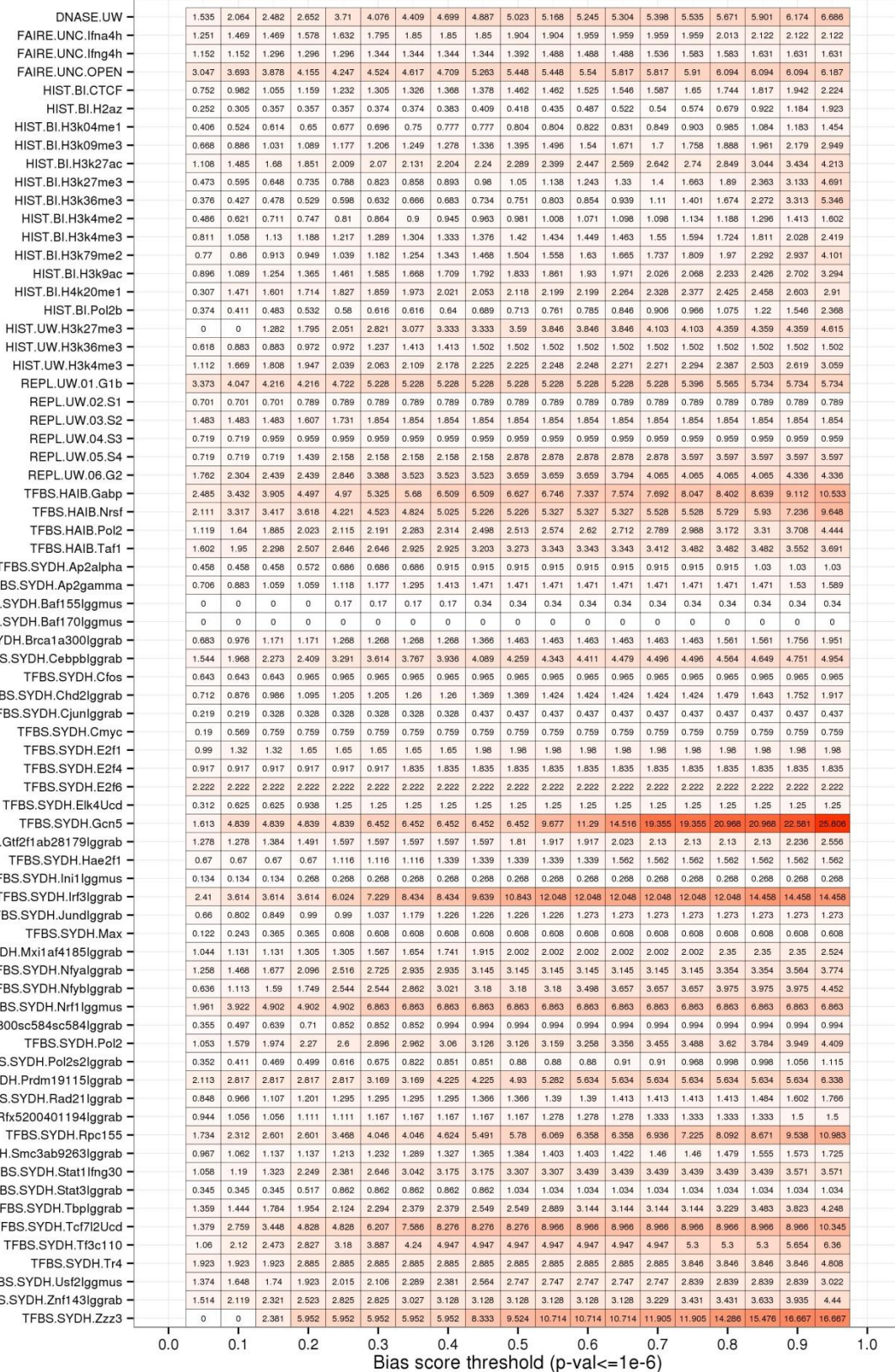


**Figure S41 | Normalized haplotype imbalance scores by copy number.**

Normalized haplotype imbalance scores were calculated and split by the underlying HRCN (total CN : hapA CN : hapB CN). The majority of the HeLa genome has a higher haplotype A copy number (as per naming conventions) and therefore expected allele balances of haplotype A over total are shifted closer to 1 (except in haplotype-balanced regions, ie 2:1:1 and 4:2:2). This results in a reduced ability to call outliers of excessive haplotype A contribution due to the reduced range of allele balance from the null hypothesis to 1 (eg. for HRCN 3:2:1, the range for haplotype B to be considered an excessive contributor is  $0.33 < B \leq 1$  whereas the range for haplotype A is  $0.66 < A \leq 1$ ).

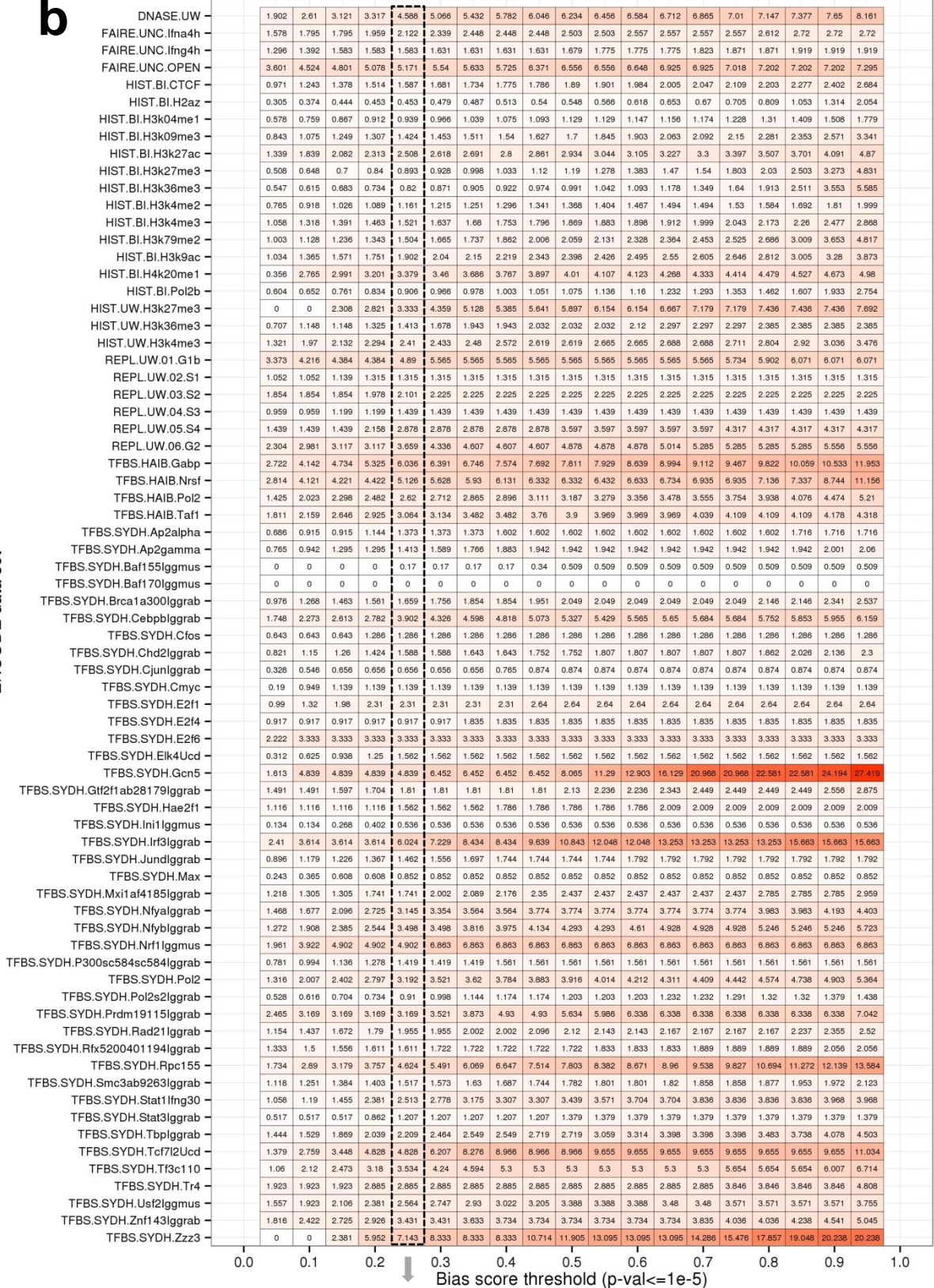


a





**b**



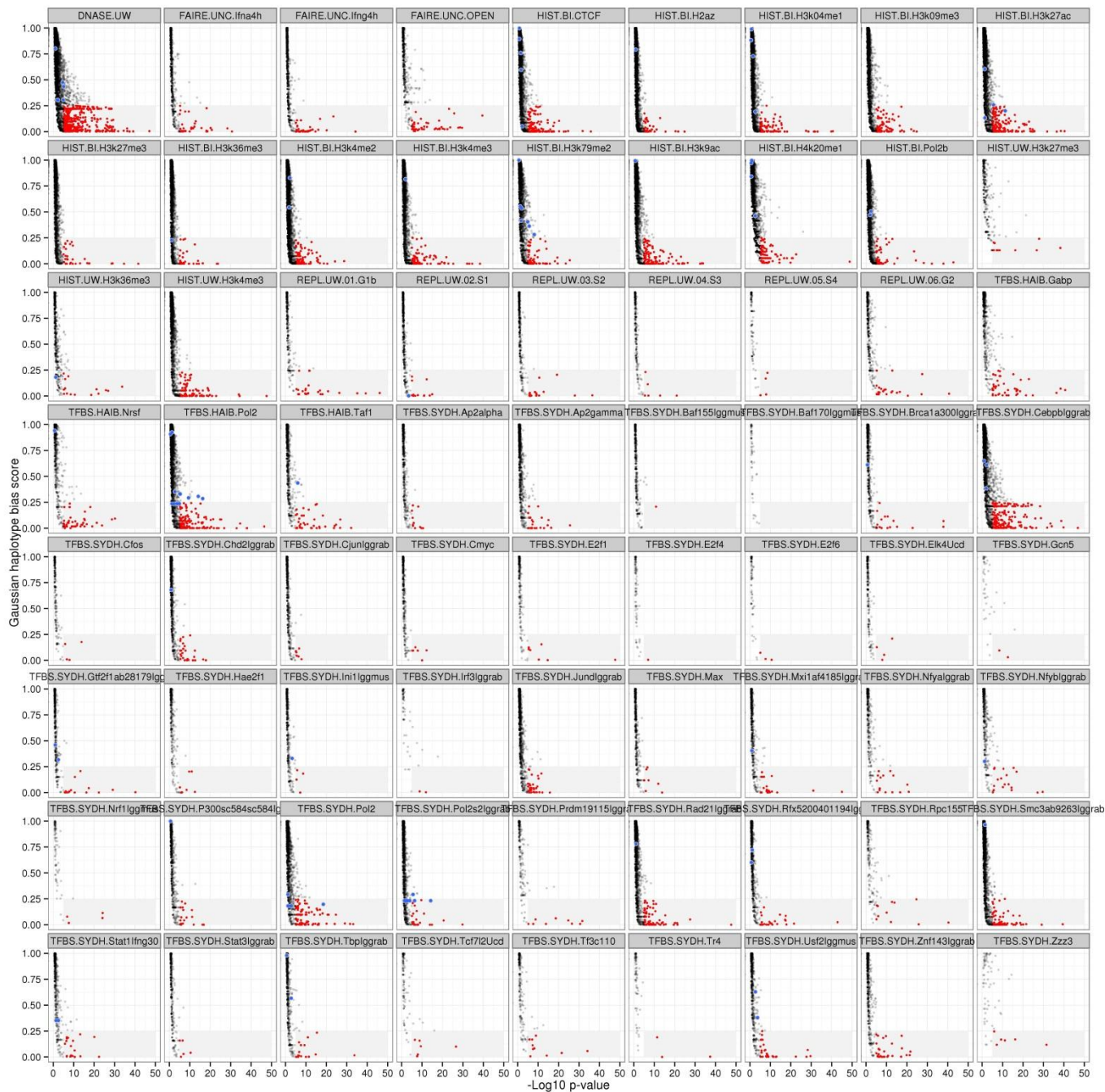


C

DNASE.UW	2.149	3.045	3.667	3.923	5.586	6.208	6.712	7.198	7.581	7.846	8.127	8.332	8.545	8.775	8.963	9.108	9.338	9.611	10.123
FAIRE.UNC.lfna4h	2.067	2.503	2.503	2.775	2.938	3.156	3.264	3.264	3.319	3.319	3.373	3.373	3.373	3.373	3.591	3.7	3.808	3.808	3.808
FAIRE.UNC.lfng4h	1.536	1.679	1.967	2.015	2.063	2.111	2.111	2.111	2.159	2.255	2.351	2.351	2.351	2.447	2.495	2.495	2.543	2.543	2.543
FAIRE.UNC.OPEN	3.693	4.894	5.448	5.725	5.817	6.556	6.646	6.741	7.064	7.849	7.941	8.033	8.495	8.587	8.68	8.864	8.864	8.864	8.957
HIST.BI.CTCTF	1.201	1.596	1.817	1.963	2.078	2.203	2.329	2.402	2.423	2.558	2.59	2.673	2.726	2.776	2.84	2.934	3.006	3.133	3.415
HIST.BI.H2az	0.505	0.609	0.687	0.696	0.705	0.731	0.766	0.792	0.818	0.835	0.862	0.922	0.957	0.975	1.009	1.114	1.358	1.619	2.358
HIST.BI.H3k04me1	0.732	0.976	1.111	1.192	1.228	1.292	1.382	1.427	1.472	1.517	1.536	1.563	1.572	1.59	1.644	1.725	1.825	1.924	2.195
HIST.BI.H3k09me3	1.148	1.438	1.656	1.743	1.903	1.976	2.063	2.092	2.208	2.281	2.44	2.542	2.775	2.804	2.862	3.021	3.094	3.312	4.082
HIST.BI.H3k27ac	1.705	2.338	2.74	3.032	3.287	3.494	3.616	3.775	3.908	4.042	4.188	4.322	4.444	4.586	4.663	4.773	4.968	5.357	6.137
HIST.BI.H3k27me3	0.735	0.98	1.05	1.208	1.26	1.295	1.4	1.47	1.593	1.68	1.803	1.908	1.995	2.065	2.328	2.556	3.028	3.796	5.356
HIST.BI.H3k36me3	0.632	0.769	0.888	0.939	1.025	1.093	1.144	1.161	1.213	1.247	1.298	1.349	1.435	1.605	1.896	2.169	2.787	3.809	5.841
HIST.BI.H3k4me2	1.125	1.359	1.539	1.674	1.801	1.927	2.017	2.062	2.134	2.17	2.233	2.332	2.368	2.377	2.422	2.476	2.584	2.701	2.89
HIST.BI.H3k4me3	1.449	1.753	1.854	1.97	2.043	2.158	2.216	2.318	2.361	2.434	2.477	2.521	2.564	2.651	2.694	2.825	2.912	3.129	3.52
HIST.BI.H3k79me2	1.397	1.665	1.916	2.059	2.31	2.543	2.722	2.919	3.116	3.206	3.331	3.582	3.671	3.886	3.994	4.155	4.477	5.122	6.286
HIST.BI.H3k9ac	1.323	1.709	2.012	2.205	2.44	2.591	2.715	2.784	2.922	3.046	3.115	3.212	3.336	3.405	3.446	3.611	3.804	4.08	4.673
HIST.BI.H4k20me1	0.386	3.686	4.365	4.931	5.174	5.497	5.853	6.063	6.451	6.726	6.855	7.001	7.211	7.276	7.518	7.583	7.783	7.828	8.036
HIST.BI.Pol2b	1.015	1.123	1.317	1.425	1.522	1.607	1.631	1.667	1.776	1.836	1.933	1.981	2.078	2.15	2.223	2.331	2.476	2.803	3.624
HIST.UW.H3k27me3	0	0	3.59	4.103	5.385	6.41	7.436	7.692	8.205	8.462	8.974	8.974	9.744	10.513	10.513	11.026	11.282	11.282	11.538
HIST.UW.H3k36me3	0.795	1.325	1.325	1.767	1.943	2.208	2.582	2.65	2.739	2.739	2.739	2.827	3.004	3.004	3.004	3.092	3.092	3.092	3.092
HIST.UW.H3k4me3	1.784	2.619	2.851	3.036	3.175	3.244	3.36	3.476	3.523	3.546	3.615	3.638	3.662	3.662	3.685	3.778	3.893	4.009	4.45
REPL.UW.01.G1b	3.273	4.216	4.384	4.384	5.228	5.902	6.071	6.239	6.408	6.408	6.408	6.408	6.408	6.408	6.408	6.408	6.408	6.408	6.408
REPL.UW.02.S1	1.227	1.315	1.402	1.578	1.753	1.753	1.84	1.84	1.84	1.84	1.928	1.928	1.928	1.928	1.928	1.928	1.928	1.928	1.928
REPL.UW.03.S2	1.978	1.978	2.101	2.349	2.596	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719	2.719
REPL.UW.04.S3	1.679	1.679	1.918	1.918	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158	2.158
REPL.UW.05.S4	2.158	2.158	2.158	2.678	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597	3.597
REPL.UW.06.G2	2.439	3.388	3.523	3.794	4.472	5.149	5.556	5.691	5.962	6.233	6.233	6.233	6.369	6.64	6.775	6.911	6.911	7.317	7.317
TFBS.HAIB.Gabp	2.84	4.379	4.97	5.799	6.509	6.864	7.219	8.166	8.521	8.757	8.994	9.822	10.178	10.296	10.789	11.124	11.479	12.071	13.491
TFBS.HAIB.Nrsf	3.518	5.025	5.226	5.427	6.432	6.935	7.337	7.538	7.94	8.342	8.543	8.744	8.945	8.945	9.246	9.548	11.156	13.568	
TFBS.HAIB.Pol2	1.701	2.421	2.743	3.003	3.218	3.371	3.57	3.647	3.877	4.03	4.183	4.306	4.459	4.551	4.796	4.949	5.087	5.486	6.221
TFBS.HAIB.Taf1	1.95	2.437	3.064	3.482	3.76	3.83	4.248	4.248	4.596	4.805	4.944	5.014	5.153	5.223	5.292	5.292	5.292	5.362	5.501
TFBS.SYDH.Ap2alpha	0.915	1.144	1.259	1.602	1.831	1.831	1.831	2.059	2.059	2.059	2.059	2.059	2.059	2.059	2.059	2.059	2.174	2.174	2.174
TFBS.SYDH.Ap2gamma	0.942	1.177	1.589	1.589	1.707	1.942	2.119	2.237	2.295	2.295	2.413	2.413	2.413	2.413	2.413	2.413	2.413	2.472	2.531
TFBS.SYDH.Baf155lggms	0	0	0.17	0.17	0.34	0.34	0.34	0.34	0.679	0.849	0.849	0.849	0.849	0.849	0.849	0.849	0.849	0.849	0.849
TFBS.SYDH.Baf170lggms	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515	1.515
TFBS.SYDH.Brca1a300lggrab	1.366	1.659	1.854	2.049	2.146	2.244	2.341	2.341	2.439	2.537	2.537	2.537	2.537	2.537	2.634	2.634	2.829	3.024	
TFBS.SYDH.Cebpblgggrab	2.257	3.088	3.563	3.902	5.327	5.921	6.21	6.617	6.905	7.211	7.397	7.635	7.788	7.855	7.906	7.991	8.093	8.195	8.398
TFBS.SYDH.Cfos	1.286	1.286	1.608	2.251	2.251	2.251	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572	2.572
TFBS.SYDH.Chd2lggrab	1.26	1.698	1.972	2.191	2.355	2.355	2.41	2.41	2.519	2.519	2.574	2.629	2.629	2.629	2.629	2.629	2.846	2.967	3.122
TFBS.SYDH.Cjunlggrab	0.656	0.984	1.202	1.202	1.311	1.311	1.311	1.421	1.53	1.53	1.53	1.53	1.53	1.53	1.53	1.53	1.53	1.53	1.53
TFBS.SYDH.Cmyc	0.19	1.139	1.328	1.328	1.328	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518	1.518
TFBS.SYDH.E2f1	1.65	1.98	2.64	3.3	3.3	3.3	3.3	3.3	3.63	3.63	3.63	3.63	3.63	3.63	3.63	3.63	3.63	3.63	3.63
TFBS.SYDH.E2f4	0.917	0.917	0.917	0.917	0.917	0.917	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835	1.835
TFBS.SYDH.E2f6	2.222	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333
TFBS.SYDH.Elk4Ucd	0.312	0.625	0.938	1.25	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562	1.562
TFBS.SYDH.Gcn5	1.613	4.839	4.839	4.839	4.839	6.452	6.452	6.452	6.452	8.065	12.903	14.516	17.742	22.581	22.581	24.194	24.194	25.806	29.432
TFBS.SYDH.Gt2f1ab28179lggrab	1.704	1.704	1.81	2.023	2.13	2.13	2.13	2.13	2.449	2.556	2.556	2.662	2.769	2.769	2.769	2.769	2.875	3.195	
TFBS.SYDH.Hae2f1	1.768	2.009	2.009	2.009	2.455	2.455	2.679	2.679	2.679	2.679	2.679	2.902	2.902	2.902	2.902	2.902	2.902	2.902	2.902
TFBS.SYDH.Ini1lggms	0.134	0.134	0.268	0.402	0.67	0.804	0.938	0.938	0.938	0.938	1.072	1.072	1.206	1.206	1.206	1.206	1.206	1.206	1.206
TFBS.SYDH.Irf3lggrab	3.614	4.819	4.819	4.819	7.229	8.434	9.639	10.843	12.048	13.253	14.458	14.458	15.663	16.867	16.867	18.072	20.482	20.482	20.482
TFBS.SYDH.Jundlggrab	1.367	1.744	1.886	2.122	2.216	2.357	2.499	2.546	2.64	2.64	2.64	2.687	2.687	2.687	2.687	2.687	2.687	2.687	2.687
TFBS.SYDH.Max	0.487	0.608	1.095	1.095	1.46	1.46	1.46	1.46	1.46	1.46	1.582	1.582	1.582	1.582	1.582	1.582	1.582	1.582	1.582
TFBS.SYDH.Mxi1af4185lggrab	1.393	1.48	1.567	2.089	2.089	2.35	2.524	2.611	2.785	2.872	2.959	2.959	2.959	2.959	2.959	3.307	3.307	3.307	3.481
TFBS.SYDH.Nfyalggrab	1.677	1.887	2.306	2.935	3.564	3.774	4.193	4.193	4.822	4.822	4.822	4.822	4.822	5.031	5.241	5.241	5.451	5.661	
TFBS.SYDH.Nfyblggrab	1.431	2.226	2.703	3.021	3.975	3.975	4.293	4.452	4.61	4.926	5.087	5.584	5.882	6.041	6.359	6.677	6.677	7.154	
TFBS.SYDH.Nrf1lggms	2.941	4.902	5.882	5.882	5.882	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824	8.824
TFBS.SYDH.P300sc584sc584lggrab	0.994	1.419	1.703	2.058	2.342	2.342	2.342	2.484	2.555	2.555	2.555	2.555	2.555	2.555	2.555	2.555	2.555	2.555	2.555
TFBS.SYDH.Pol2	1.415	2.369	2.929	3.488	3.962	4.442	4.574	4.804	4.936	5.002	5.133	5.462	5.627	5.756	5.791	5.923	6.088	6.252	6.713
TFBS.SYDH.Pol2s2lggrab	0.704	0.822	0.939	0.996	1.262	1.438	1.585	1.614	1.614	1.673	1.702	1.731	1.761	1.79	1.849	1.878	1.878	1.937	1.995
TFBS.SYDH.Prdm19115lggrab	2.817	3.521	3.521	3.521	3.521	3.873	4.225	5.634	5.986	6.69	7.042	7.394	7.394	7.394	7.394	7.746	7.746	7.746	8.451
TFBS.SYDH.Rad21lggrab	1.625	2.00																	

**Figure S42 | Haplotype imbalanced ENCODE peak percentages.**

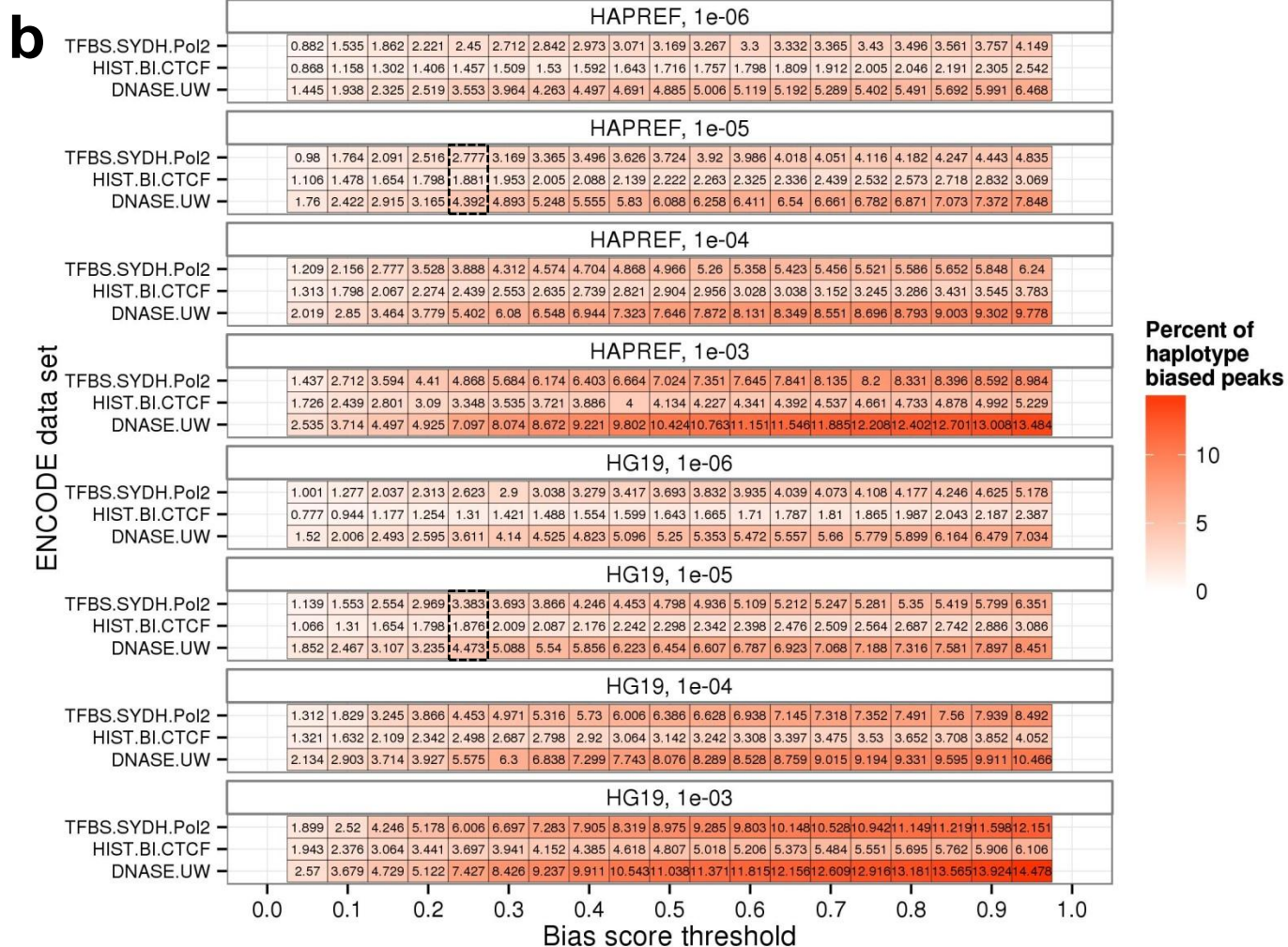
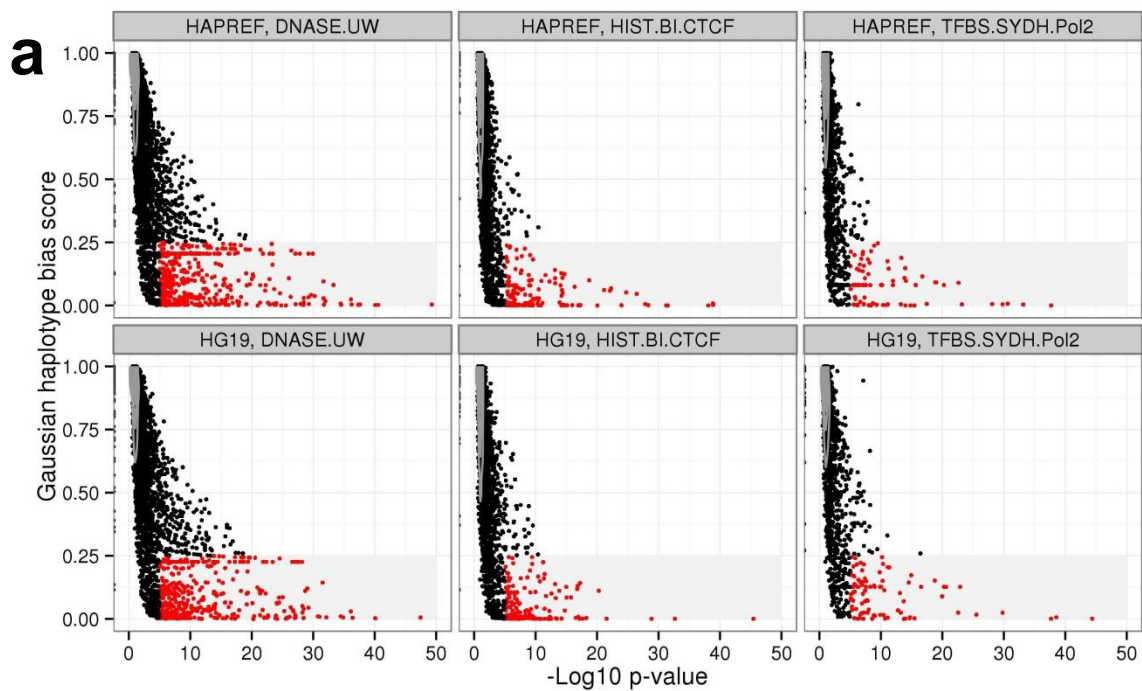
Percentages of peaks within each ENCODE enrichment data set called as outliers at three thresholds (**a.**  $P < 1e-6$ , **b.**  $P < 1e-5$ , and **c.**  $P < 1e-4$ ) with respect to normalized Gaussian haplotype imbalance score. The dashed box in **b.** represents the scoring threshold used of a p-value of  $1e-5$  and normalized imbalance score of 0.25.



**Figure S43 | ENCODE peak haplotype imbalance scoring.**

For each peak with an ENCODE data track, the normalized haplotype imbalance score is plotted against the  $-\log_{10}$  p-value (the degree of significance against the null hypothesis of haplotype-balanced signal). Gray boxes with red points represent peaks called as outliers at a  $P < 1e-5$  and normalized haplotype imbalance score of  $\leq 0.25$ . Blue dots represent peaks near the HPV-18 / MYC locus.





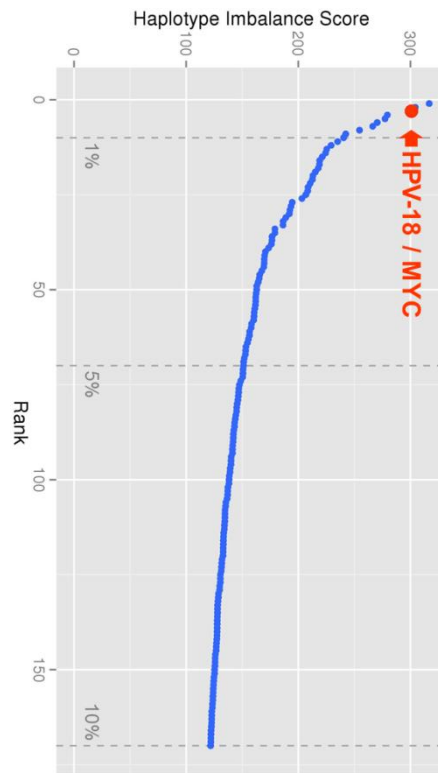
**Figure S44 | ENCODE peak reference bias effects on outlier calling.**

The use of a HeLa-specific haplotype-resolved reference eliminates the reference bias, but does not substantially change the set of peaks called as outliers. **a.** Haplotype imbalance scores when aligning to a haplotype-resolved HeLa reference (top) or hg19 (bottom). **b.** Percentage of peaks called as outliers with  $P < 1e-5$  and an imbalance score cutoff of 0.25. Using HeLa haplotype-specific reference sequences changes the set of outliers called by only 0.606% 0.005%, and 0.081% for Pol2 ChIP-seq, CTCF ChIP-seq, and DNaseHS-seq, respectively.

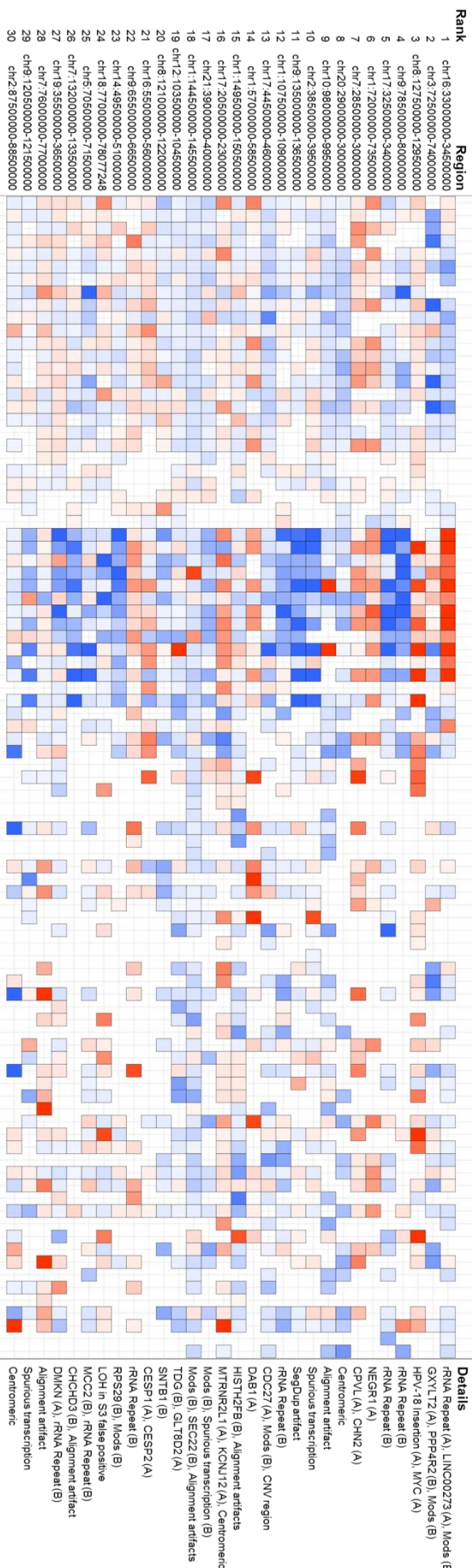
## a ENCODE Outlier Window Scores (All)



## b ENCODE Outlier Window Scores (Top 10% of Data)



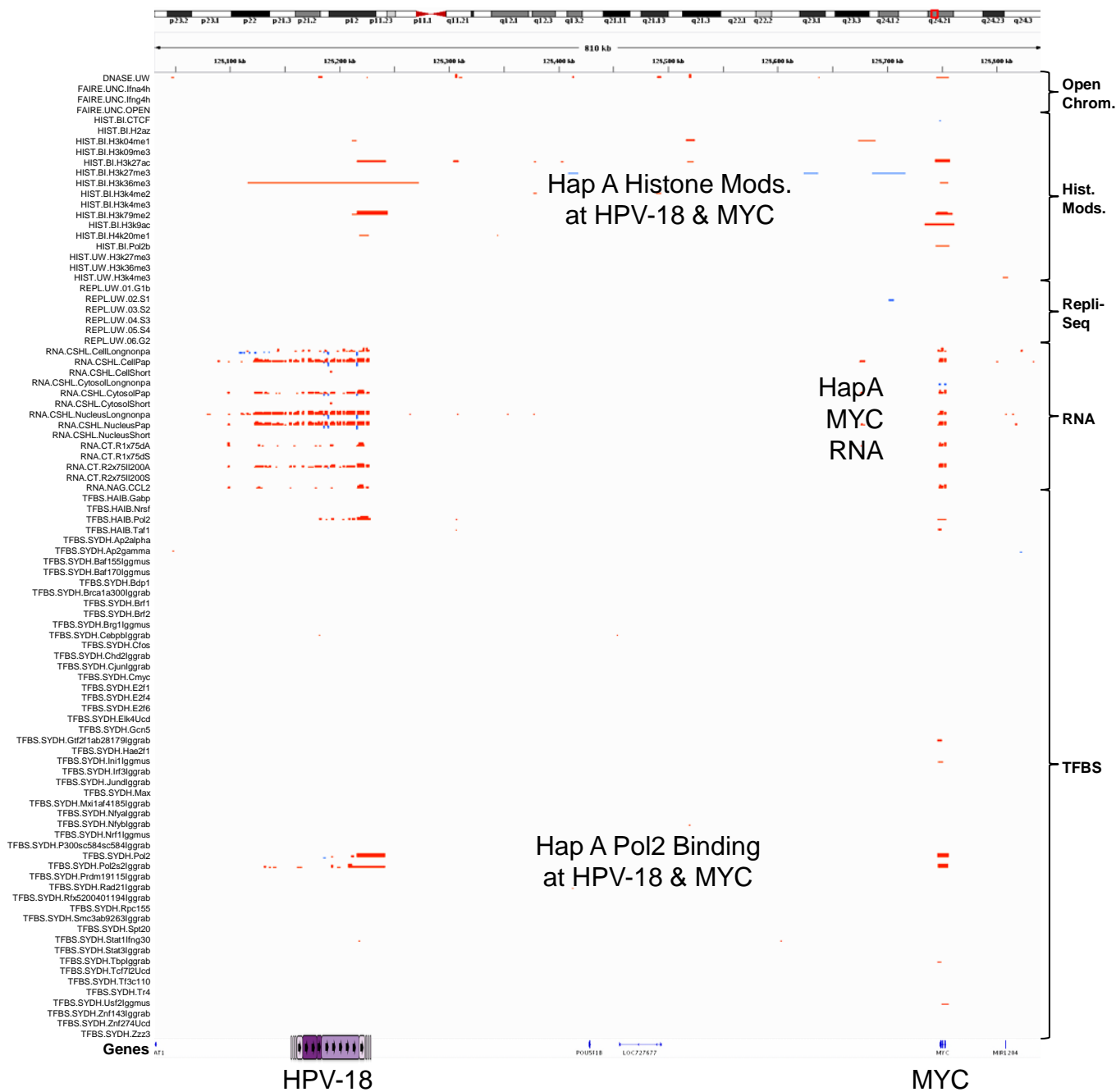
## ENCODE Top 30 Haplotype Outliers



**Figure S45 | Phased ENCODE outlier analysis.**

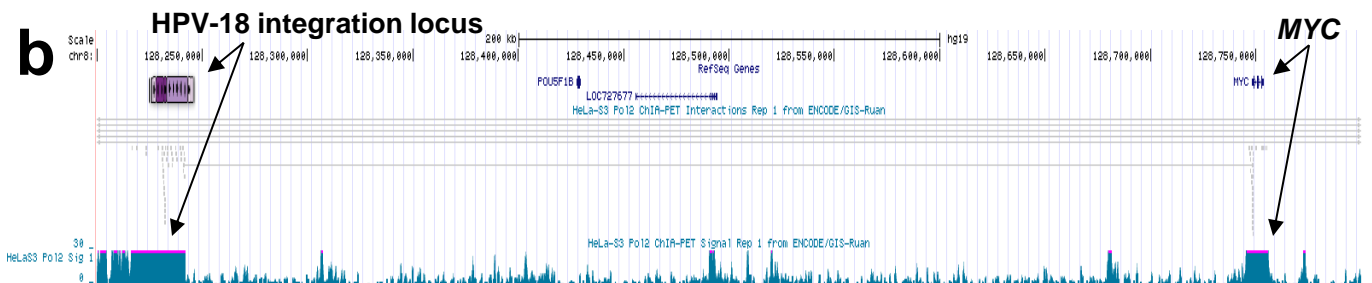
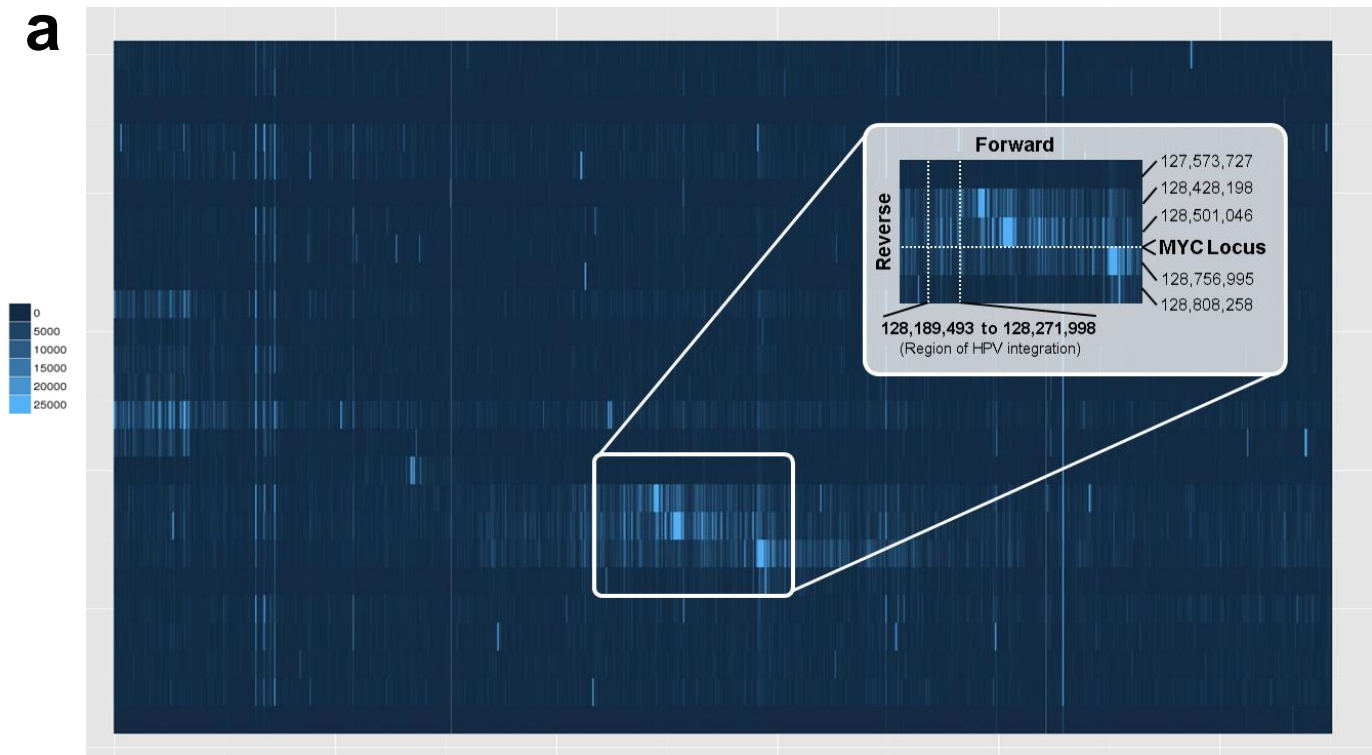
**a.** Ranked sliding window scores for ENCODE outliers. The window containing the HPV-18 integration and MYC loci is highlighted in red. **b.** Closer investigation of the top 50 scores from **a.** **c.** Top 30 outlying windows after being condensed for overlapping regions. Columns represent haplotype imbalance scores for haplotype A (red) and haplotype B (blue) for the ENCODE tracks across each region.





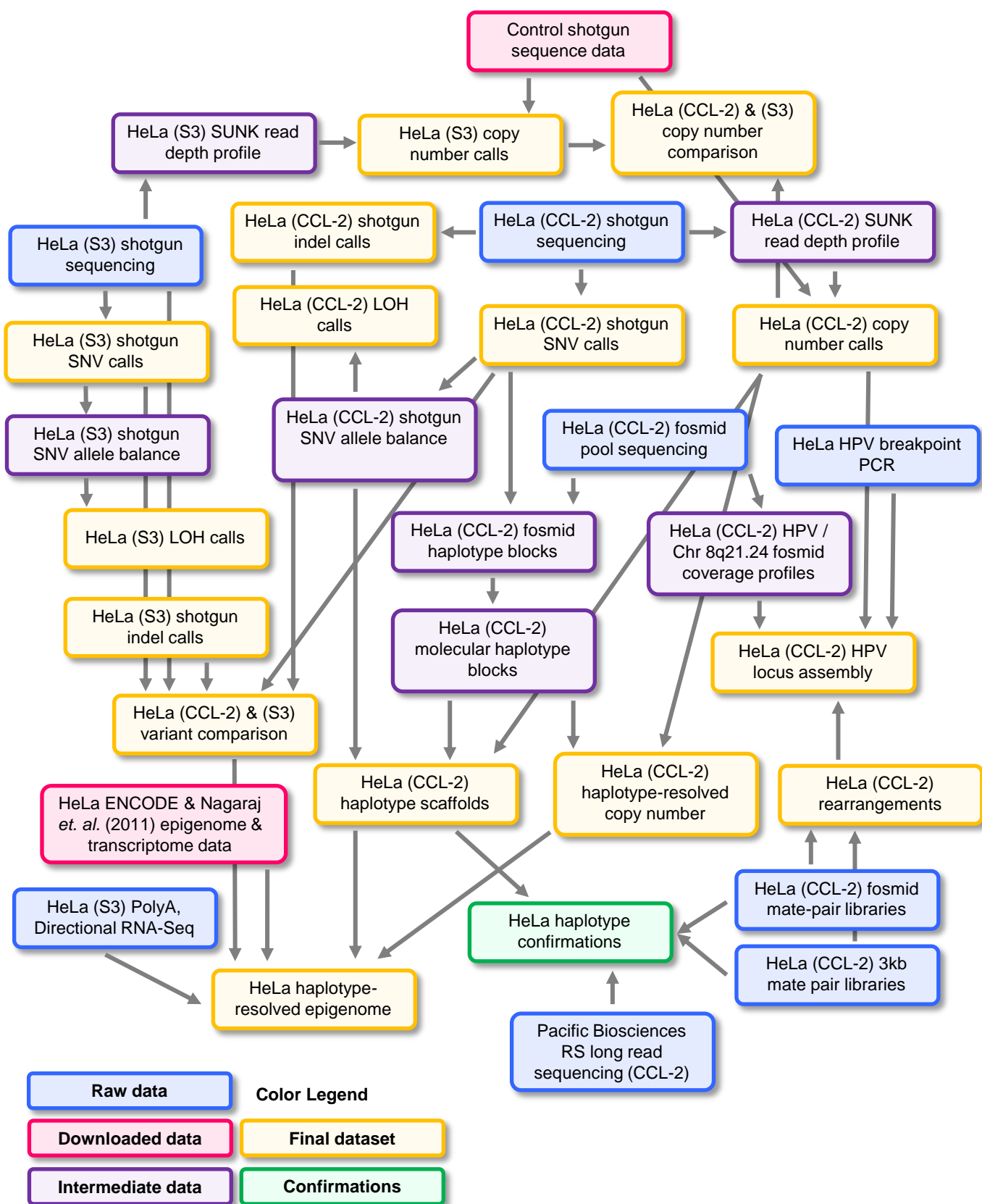
**Figure S46 | ENCODE haplotype imbalances for HPV-18 and MYC.**

Red peaks indicate haplotype A imbalance, blue peaks represent haplotype B for copy number normalized haplotype imbalance scores.



**Figure S47 | Long range with *MYC* from 5C and ChIA-PET data.**

**a.** ENCODE 5C chromatin interaction data (available only for the GM12878 cell line) demonstrates long-range interactions between *MYC* and distal upstream sites. The highlighted region includes the site of HPV-18 integration (into the HeLa but not GM12878 genome). **b.** Spanning reads from ENCODE ChIA-PET data in HeLa S3 cells indicate long range integration between the HPV-18 interaction and site and *MYC* locus. Teal profile represents Pol2 signal and contains peaks at the HPV-18 and *MYC* loci.



**Figure S48 | Datasets and analyses for HeLa CCL-2 and HeLa S3.**